Optimized GeLC-MS/MS for Bottom-Up Proteomics

DISSERTATION

zur Erlangung des akademischen grades
Doctor rerum naturalium

(Dr. rer. nat.)

vorgelegt der

Fakultät Mathematik und Naturwissenschaften

der Technischen Universität Dresden

von

Dipl. Chemie - Ingenieurin (FH)

Natalie Wielsch (geborene Schmalz)

Geboren am 20. Juni 1973 in Frunse / Kirgisien

Gutachter:

Professor Dr. Michael Göttfert, Technische Universität Dresden

Dr. Christoph Thiele, Max Planck Institut für Molekulare Zellbiologie und Genetik, Dresden

Professor Dr. Marek Šebela, Palacký University, Olomouc, Czech Republic

Tag der Einreichung:	18.12.2008
Tag der Verteidigung:	14.05.2009

"I want to know how God created this world. I am not interested in this or that phenomenon, in the spectrum of this or that element. I want to know His thoughts; the rest are details."

Albert Einstein

TABLE OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

ABBREVIATIONS

ACD-BT	α-cyclodextrin modified bovine trypsin
ADP	accelerated digestion protocol
BAPNA	Nα-benzoyl-DL-argenine 4-nitroanilide
BCD-BT	β-cyclodextrin modified bovine trypsin
BLAST	basic local alignment search tool
BT	bovine trypsin
CID	collision-induced dissociation
CDP	conventional digestion protocol
DB	database
E-value	expectation value
ECD	electron capture dissociation
ESI	electrospray ionisation
ETD	electron transfer dissociation
FWHM	full width at half maximum
FT	Fourier transform
FT-ICR MS or FTMS	Fourier transform-ion cyclotron resonance mass spectrometry
HPLC	high performance liquid chromatography
HSP	high scoring segment pair
ICAT	isotope coded affinity tags
IEF	isoelectric focusing
iTRAQ	isotope tags for relative and absolute quantification
LAC-BT	lactose modified bovine trypsin
LIT	linear ion trap
MAL-BT	maltose modified bovine trypsin
MALDI	matrix-assisted laser desorption ionisation
MAT-BT	maltotriose modified bovine trypsin
MEL-BT	melibiose modified bovine trypsin
MET-PT	methylated porcine trypsin
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MS BLAST	mass spectrometry driven BLAST
mRNA	messenger ribonucleic acid
MRM	multiple reaction monitoring
m/z	mass-to-charge ratio
nanoESI	nanoelectrospray
nanoLC-MS/MS	nanoflow liquid-chromatography-tandem MS
PAGE	polyacrylamide gel electrophoresis
PMF	peptide mass fingerprinting
PTM	post-translational modifiecations
Q(q)TOF	quadrupole time-of-flight mass spectrometer
RAF-BT	raffinose modified bovine trypsin
RAFR-BT	RAF plus biacetyl
SDS	sodium dodecyl sulfate
SILAC	stable isotope labeling with amino acids in cell culture
STA-BT	stachyose modified bovine trypsin
TOF	time-of-flight

ACKNOWLEDGEMENTS

First I would like to thank my supervisor Dr. Andrej Shevchenko and within International PhD program, and especially Prof. Dr. Wieland Huttner, Dr. Birgit Knepper-Nicolai and Jana Marschner for inviting me to the Max Plank Institute of Molecular Cell Biology and Genetics in Dresden. Everyone at MPI-CBG has been a friend and colleague and made my time in Dresden fascinating and exciting. I am especially grateful to librarians and my friends Silke Thüm, Anke Kahnert, Carola Schuchardt for providing great library resources and for their assistance and encouragement in difficult times.

I would like to thank the members of my thesis advisory committee: Prof. Dr. Michael Göttfert and Dr. Christoph Thiele for their guidance during this thesis work, valuable discussions and for a large number of excellent suggestions on this topic. I am indebted to my supervisor Dr. Andrej Shevchenko for his support through my research and for supplying a great working environmental and all the resources I needed.

All the members of the mass spectrometry laboratory have contributed to the work in some way: Dr. Henrik Thomas, Dr. Patrice Waridel, Vineeth Surendranath, Dr. Vinzenz Link, Andrea Knaust, Dr. Anna Shevchenko, Magno Junqueira, Dr. Dominik Schwudke, Dr. Christer Ejsing, Dr. Vitaly Matiash, Kai Schuhmann and Dr. Marina Edelson-Averbukh. I want to thank them for all their help, support, interest and valuable hints.

I would like to thank Prof. Dr. Marek Šebela and members of his laboratory who made a direct contribution to the ideas and data contained in the present work and provided me with supporting materials. I thank Prof. Dr. Jan Havliš for his help and advice by application of ^{18}O labeling quantification method. I am especially grateful to Dr. Nurhan Özlu, Dr. Anne-Lore Schlaitz and Prof. Dr. Tony Hyman for long-standing collaboration.

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. Especially, I would like to give my special thanks to my beloved parents whose patient love, encouragement and support enabled me to complete this work. I also thank my beloved husband Torsten for daily supporting me through this work.

SUMMARY

Despite tremendous advances in mass spectrometry instrumentation and mass spectrometry-based methodologies, global protein profiling of organellar, cellular, tissue and body fluid proteomes in different organisms remains a challenging task due to the complexity of the samples and the wide dynamic range of protein concentrations. In addition, large amounts of produced data make result exploitation difficult. To overcome these issues, further advances in sample preparation, mass spectrometry instrumentation as well as data processing and data analysis are required.

The study presented here focuses as first on the improvement of the proteolytic digestion of proteins in gel based proteomic approach (Gel-LCMS). To this end commonly used bovine trypsin (BT) was modified with oligosaccharides in order to overcome its main disadvantageous, such as weak thermostability and fast autolysis at basic pH. Glycosylated trypsin derivates maintained their cleavage specifity and showed better thermostability, autolysis resistance and less autolytic background than unmodified BT. Trypsins conjugated with maltotriose (MAT-BT), raffinose (RAF-BT) and RAF-BT with additionally modified by biacetyl arginine residues (RAFR-BT) were considered as perspective candidates for gel-based proteomics applications [1]. In line with the "accelerated digestion protocol" (ADP) previously established in our laboratory [2] modified enzymes were tested in in-gel digestion of proteins. Kinetics of in-gel digestion was studied by MALDI TOF mass spectrometry using ^{18}O-labeled peptides as internal standards as well as by label-free quantification approach, which utilizes intensities of peptide ions detected by nanoLC-MS/MS. In the performed kinetic study I characterized the effect of temperature, enzyme concentration and digestion time on the yield of digestion products. The obtained results showed that in-gel digestion of proteins by glycosylated trypsin conjugates was less efficient compared to the conventional digestion (CD) and achieved maximal 50 to 70% of CD yield, suggesting that the attached sugar molecules limit free diffusion of the modified trypsins into the polyacrylamide gel pores. Nevertheless, these thermostable and autolysis resistant enzymes can be regarded as promising candidates for gel-free shotgun approach.

To address the reliability issue of proteomic data I further focused on protein identifications with borderline statistical confidence produced by database searching. These hits are typically produced by matching a few marginal quality MS/MS spectra to database peptide sequences and represent a significant bottleneck in proteomics. A method was developed for rapid validation of borderline hits, which takes advantage of the independent interpretation of the acquired tandem mass spectra by *de novo* sequencing software PepNovo followed by mass-spectrometry driven BLAST (MS BLAST) sequence similarity searching that utilize all partially accurate, degenerate and redundant proposed peptide sequences [3]. This validation approach was applied in two collaboration projects, which aimed to study centrosomal effectors of *C.elegans* mitotic spindle assembly. In the first study, which aimed to determine interaction partners of the protein TPXL-1 [4], about 300 proteins were identified by nanoLC -MS/MS analysis and database searching, more than 50% of them were of borderline statistical confidence. PepNovo/MS BLAST enabled rapid assignment (confirmation or rejection) of more than 70% of these hits. In the second study, PepNovo/MS BLAST was applied for validation of ambiguous hits obtained by identification of proteins associated with the novel protein RSA-1 (RSA complex) [5].

1 INTRODUCTION

1.1 From genomics to proteomics

The study of an organism's genome is fundamental for understanding its biology [6] Advances in this field, such as improvements in DNA sequencing, bioinformatics and application of microarray technology to characterize gene expression profiles demonstrated the power of high throughput and enabled understanding how genes are organized and regulated [7-9]. Complete genomic sequences for different organisms were provided [10-12], including entire human genome [13-15]. Although the number of genes is relatively small and ranges from a few hundred for bacteria to tens of thousands for mammalian species, prediction of possible expressed proteins is a complex task, since the same gene can produce multiple protein products by alternative splicing of pre-mRNA transcripts or different post-translational modifications (PTM) of expressed proteins. Thus, the number of proteins in a species proteome exceeds by far the number of genes in the corresponding genome. Genomic approaches also cannot predict where proteins are localized in a cell and in what quantity and molecular form they are present [16].

Proteins are involved in all biological processes and considered as most important biological molecules. They are characterized by their amino acid sequence, relative expression (measured in copies per cell), specific activity, state of modification and association with other proteins or different biological molecules Figure (1.2). The systematic analysis of protein complement expressed by a genome has been named proteomics [17]. Proteome reflects the cellular state in dependence of the physiological conditions and is highly dynamic: expressed proteins differ in their abundance, state of modification and subcellular location [18]. Therefore the crucial goal of proteomics research is directed toward the systematic study of diverse cellular states and molecular mechanisms which control them, so providing to understanding of fundamental biological processes[19]. In other words proteomics aims to explain the information contained in a genome in terms of the structure and biological function [20].

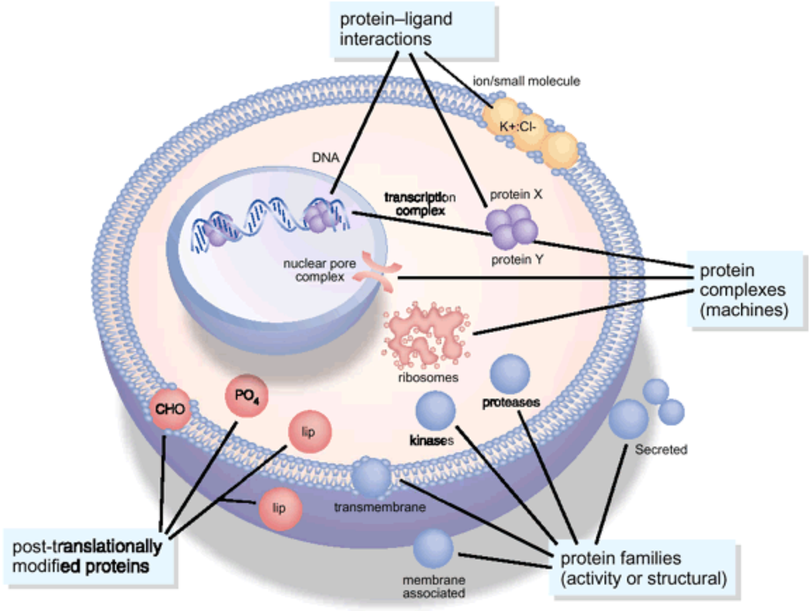


Figure 1.1 Representation of a eukaryotic cell.

A section through eukaryotic cell highlights diverse properties of proteins: the subcellular distribution, quantity, modification and interaction state, catalytic activity and structure (adapted from [17]).

1.2 Mass spectrometry based proteomics

1.2.1 Ionization techniques

For long time mass spectrometry was mostly applied for the analysis of small and thermostable compounds because of lack of effective techniques to and transfer the ionized molecules from the condensed phase into the gas phase without excessive fragmentation and then softly ionize to yield the intact molecular ion [21].

The development of two major soft ionization techniques electrospray ionisation (ESI) [22] and matrix-assisted laser desorption/ionisation (MALDI) [23] enabled application of mass spectrometry for generating ions from large, nonvolative analytes such as proteins and peptides. Those techniques are complementary and differ in way how molecules are converted into ions:

1) In ESI, charged droplets are produced by passing a solubilised sample through a high voltage needle at atmospheric pressure, followed by their desolvation (till analyte is a solvent-free molecular beam) prior to entrance into the high vacuum of the mass

spectrometer; this ionization technique allows on-line coupling to chromatography or electrophoresis.

2) In MALDI samples are cocrystallized on a sample plate with a small organic matrix compound that usually has an aromatic ring structure, which absorbs at the wavelength of the laser; the analyte is then ablated and ionized out of dry, crystalline matrix via laser pulses.

MALDI and ESI differ in charge of produced ions; MALDI ionization results predominantly in single charged ions, in ESI MS tryptic peptides are typically ionized as doubly or triply charged ions. Multiply charged ions can be efficiently fragmented at lower collision energy in contrast to single charged ions, which require higher collision energy. It is also known that differences in ionization efficiencies exist between these two ionization methods. For instance, positive ion ESI preferentially ionizes hydrophobic peptides [24] while MALDI has been reported to preferentially ionize basic [25] and aromatic residues [26]. In general, MALDI is more tolerant to salts and buffer components [18] while the determination of low-mass peptides might be better done by ESI than MALDI because of chemical noise associated with MALDI matrix peaks [27]. Both ionization techniques provide to some extent complementary information, allowing their use in combination to maximize protein sequence coverage [26, 28, 29].

1.2.2 MS instrumentation

On the basis of ESI and MALDI ionization techniques different mass spectrometers were developed to address various proteomics questions. Generally, mass spectrometers measure the mass-to-charge ratio of analytes; for protein studies this can include intact proteins and protein complexes, fragment ions produced by gas-phase activation of protein ions (top-down sequencing), peptide produced by enzymatic or chemical digestion of proteins (mass mapping), and fragment ions produced by gas-phase activation of mass-selected peptide ions (tandem mass spectrometry) [30]. Mass spectrometers consist of three basic components: an ion source, a mass analyser and an ion detector. The mass analyser is central to the technology. In the context of proteomics, its key parameters are resolution, sensitivity, mass accuracy and the ability to generate information-rich ion mass spectra from peptide fragments [31]. There are several basic types of mass analysers: quadrupole, time-of-flight, ion trap, Fourier transform (FT) and Orbitrap.

Quadrupole mass analyzer developed by Wolfgang Paul consists of 4 circular rods, set parallel to each other. It works as a filter, which selectively isolates sample ions based on the stability of their trajectories in the oscillating electric fields that are applied to the rods (each opposing rod pair is connected together electrically and fixed DC (direct current) and alternating RF (radio frequency) potentials applied between one pair of rods, and the other). This allows selection of an ion with particular *m/z*, or scanning a range of *m/z*-values by continuously varying the voltages.

The triple quadrupole spectrometer is one of the most popular instruments based on quadrupole mass analyzer [32]. In this device, the first Q_1 and the third Q_3 quadrupole are mass filters, in which Q_1 serves to select the precursor ion and Q_3 scans the masses of fragment ions. The fragment ions are produced in the collision cell under collision-induced dissociation (CID) enclosed in a quadrupole ion guide q_2.

A complete mass spectrum can be obtained from a quadrupole mass filter only by scanning. This considerably reduces the sensitivity of the acquisition, since different ions in the spectrum are examined one at time, discarding all others. Serious drawback of the quadrupole mass analyzer is also its low mass resolution.

In Time-of-flight (TOF) analyzer the mass-to-charge ratio of an analyte ion is computed from its flight time through a vacuum tube of the fixed length; the flight time is proportional to the square root of the m/z [33]. It is a non-scanning analyzer and is widely used in mass spectrometry because of its speed, high sensitivity and wide detectable mass range. Commercial TOF instruments can typically achieve resolution up to 40,000 (full width at half maximum, FWHM) [34]. Thus, by proper mass calibration mass accuracy in the low-parts per million (ppm) range is achievable. MALDI-TOF, Q(q)TOF and TOF-TOF are instruments based on TOF mass analyzer.

MALDI-TOF instruments operate with MALDI source and are used in analysis of intact peptides. This approach is defined as peptide mass fingerprinting (PMF) and has been proved to be powerful proteomic tool because of its characteristics of speed, robustness, sensitivity and automation [35-37]. MALDI-TOF mass spectrometers equipped with reflectrons are able to analyze fragment ions produced from precursor ions that spontaneously decompose in flight. Such ions are generally referred to as metastable ions, and the process of decomposition in the field free region between the ion source and the reflectron is commonly referred to as post source decay (PSD) [38, 39]. The analysis of such PSD ions is an established technique that is capable of providing complementary MS/MS information. However, acquisition of PSD is rather slow and less sensitive than peptide mass fingerprinting. Moreover, the spectra show low resolution and mass accuracy. Several developments such as LIFT method [40] or new "parallel PSD" technique [41] considerably reduce the analysis time of MALDI PSD spectra.

Vestal et al. [42] developed a tandem TOF mass spectrometer (MALDI-TOF/TOF) to use the high-speed capabilities of the TOF mass analyzer to create a high-throughput tandem mass spectrometer. The first TOF mass analyzer is used in the ion selection process, and the selected ions are then transferred into a collision cell. Analysis of products is performed in a second TOF mass analyzer. MALDI-TOF/TOF instruments allow high sensitive peptide analysis and comprehensive fragmentation information, using high energy collision-induced dissociation (CID) instead of relying on post source decay [43].

The Q(q)TOF mass spectrometers (referred as hybrid instruments) combine the ion selection and tandem MS capabilities of a triple quadrupole with the resolution of TOF spectrometers [44]. The quadrupole operate as ion guides in MS mode to transmit the ions to the TOF analyzer. In the MS/MS mode, the precursor ions are selected in the first quadrupole and undergo fragmentation through collision-induced dissociation in the second quadrupole (RF-only) and the product ions are subsequently analyzed in the TOF analyzer. Obtained spectra show good mass accuracy and high resolution, which allows the determination of the charge state and unambiguous assignment of the mono-isotopic masses. Q(q)TOF mass analyzers take advantage of implementation of both ionization techniques ESI and MALDI (using rapidly switchable ESI/MALDI ion source) [45, 46], which produce different data sets and complement each other [29].

Ion trap (IT) analyzer focus ions into a small volume with an oscillating electric field; ions are resonantly activated and ejected by electronic manipulations of this field. Ion traps are very sensitive, because they concentrate ions in the trapping field for varying lengths of time [47]. IT instruments allow fast data acquisition, because they can rapidly shift between MS and MS/MS modes during data collection and enable in conjunction with data-dependent experiment high-throughput analyses. However, IT analyzers have limited resolution, low ion-trapping capacity, and space-charge effects that negatively impact mass measurements accuracy [48]. The development of linear ion trap analyzer with higher ion-trapping capacities has expanded the dynamic range and the overall sensitivity of this technique [47, 49, 50]. Typically, LIT instruments have multiple-stage sequential MS/MS capabilities, often referred as MS^n in which fragment ions are iteratively isolated and further fragmented, a strategy that has proven to be very useful for the analysis of posttranslational modifications (PTM) such as phosphorylation [51].

Linear ion traps [52] can be combined with two quadrupoles (Q-Q-LIT) to create a configuration similar to a triple quadrupole. When quadrupoles are combined with an ion trap, ions can be isolated and fragmented outside the ion trap and then accumulated in the trap for analysis of the fragment ions [53]. Additionally, ions can be simply passed through the mass filters and accumulated in the linear ion trap for analysis. Q-Q-LIT instruments offer increased sensitivity and some additional features derived from quadrupole technology such as 1) precursor ion scanning, which is typically used to detect subsets of peptides in a sample that contain a specific functional group, for instance a phosphate ester or a carbohydrate modification, 2) neutral loss scanning, which is used to detect those peptides in a sample that contain a specific functional group (for instance for detection of peptides phosphorylated at serine or threonine residues via a loss of phosphoric acid), and 3) multiple reactions monitoring (MRM), which is used for the detection of a specific analyte with known fragmentation properties [53, 54].

A mass spectrometer with excellent resolving power and mass accuracy is the **Fourier transform–ion cyclotron resonance** (FT-ICR) [55, 56]. Mass measurement accuracies of 1-2 ppm and resolution in excess of 10^5 can be achieved by this instrument[57]. FT-ICR MS use high magnetic fields to trap the ions and cyclotron resonance to detect and excite the ions. An external LIT combined with FT-ICR allows isolation and fragmentation of ions outside FTMS device and so combines rapid acquisition of low-resolution MS/MS spectra with accurate measurement of precursor masses [58]. FTMS is applied in shotgun proteomics and the analysis of fragments of intact proteins, termed top-down proteomics [59]. A limitation of the hybrid ion trap FT system is the relatively slow acquisition rate (several s per cycle) and the limited dynamic range of IT devices. Another limitation of FT and hybrid FT systems is significant maintenance cost of high magnetic field detector.

A relatively new **mass analyzer called orbitrap** [60, 61] is an ion trap, which is based on a new physical principle - the ions are separated by their oscillating in an electrostatic field [62]. This instrument offers also excellent resolution and mass accuracy [63] similar to an FT-ICR mass spectrometer but without an expensive superconducting magnet. On the basis of this analyzer a new hybrid mass spectrometer was developed which combines a linear ion trap mass spectrometer and an orbitrap mass analyzer; C-shaped storage trap is used to store and collisionally cool ions before injection into the orbitrap [64]. As in hybrid FT-ICR mass spectrometers this instrument combines two mass analyzers: the fast and sensitive LIT and the orbitrap with high resolving power and mass accuracy [63, 65]. This allows experiments in which both mass analyzers work in parallel in acquiring high resolution/mass accuracy spectra of precursor ions and their fragmentation in fast linear ion trap [64, 66, 67]. Further, high mass accuracy and resolving power of this instrument allows its application in the analysis of PTM [68] and in the top-down approach, which analyzes intact proteins [69].

Despite variety of mass spectrometers no one instrument has all the features which allow ideal proteomics analysis [21]. Choice of the mass spectrometric method always depends on the analytical problem to be solved and the experimental setup.

1.3 Proteomics strategies: top-down versus bottom-up

Profiling of proteins represents a complex analytical task, because of high complexity and dynamic range of proteome. Protein abundances in a proteome ranges from five to six orders of magnitude for yeast cells and more than ten orders of magnitude for human blood serum [70]; this dynamic range exceeds the dynamic range of any analytical method or instrument [71]. To overcome this problem several separation methods were developed, which are based on physical or chemical properties of peptides/proteins, such as solubility, localization, charge, size, hydrophobicity and affinity to certain matrices [72-81]. There are, for instance, fractionation methods (differential extraction, centrifugation), chromatography (affinity, ion exchange, hydrophobic, gel filtration) and electrophoresis (1D, 2D, capillary electrophoresis) [82, 83].

There are two mass spectrometry based strategies to profile proteins: top-down proteomics, which involves direct protein fragmentation in the gas phase and bottom-up proteomics, which relies on peptide analysis of proteolyzed proteins [84]. Application of top-down or bottom-up approaches in proteomics analysis depends on the question to be answered. Given the complementary nature of the information provided by top-down and bottom-up strategies, both will continue to be employed in proteomics.

1.3.1 Top-down proteomics

In top-down approach intact proteins are ionized by ESI and subsequently fragmented in the mass spectrometer. Sufficient number of fragments provides comprehensive information of the analyzed protein and its modifications [85]. However, gas-phase fragmentation of intact protein ions, especially from large proteins has been critical aspect in bottom-up approach. Han et al. [86] demonstrated informative fragmentation of intact proteins with molecular masses exceeding 200 kDa. Significant improvement was achieved by the development of new fragmentation methods such as electron capture dissociation (ECD) [87] and electron transfer dissociation (ETD) [88, 89].

The main advantage of top-down approach (compared to peptide based strategy) is high sequence coverage up to 100% and therefore the ability to characterize all PTMs and changes in protein sequences [90]. In addition, the time-consuming protein digestion required for bottom-up methods is eliminated. The analysis of intact proteins generally requires high resolution mass measurements to resolve highly charged ions and their isotopes and has been generally performed on FT-ICR instruments [59, 91-94]. Recently Macek et al. [69] showed the applicability of LTQ-Orbitrap for top-down analysis. Further, Waanders et al. [95] extended this work by application of SILAC technology to quantification of intact proteins.

Top-down proteomics suffers from several limitations. First, separation of proteins in complex mixtures prior mass spectrometric analysis is challenging because of different physico-chemical properties of proteins [85]. Second, it is still difficult to obtain sufficient fragmentation of intact proteins larger than 50 kDa. Third, ECT and ETD offer not sufficient fragmentation efficiency, requiring long ion accumulation, activation, and detection times. Fours, there is necessary to understand comprehensively the protein dissociation mechanisms [96], including the impact of precursor ion charge state and the role of protein primary, secondary and tertiary structure, what will provide development of sophisticated bioinformatics tools [97-100]. Because of mentioned limitations application of top-down approach is restricted for special cases (analysis of PTM's) and is not used in high-throughput proteomics.

1.3.2 Bottom-up proteomics

Bottom-up proteomics relies on mass spectrometric analysis of peptides in proteolytic digests of analyzed proteins. Generally, there are two approaches for protein identification: peptide mass fingerprinting (PMF) and tandem mass spectrometry (MS/MS).

In PMF usually acquired by MALDI-TOF MS, a unique mass fingerprint of a protein is created. Because mass mapping requires an essentially purified target protein, the technique is commonly used in conjunction with prior protein fractionation using two-dimensional gel electrophoresis (2DE), where proteins are separated on the basis of their isoelectric point in the first dimension and by their molecular mass in the second dimension [35, 36]. 2DE offers several advantages: 1) ability to separate related protein forms, such as differently modified forms, 2) low sample complexity and 3) additional information obtained from 2D gel (isoelectric point and molecular mass) [101, 102]. However it has limitations when dealing with very large or small proteins, proteins at the extremes of the pI scale, membranes, and low-abundant proteins [103, 104].

Tandem mass spectrometry is more prominent technique in bottom-up proteomics, since it elucidates structural features of the analysed peptides. Generally, there are two main approaches to analyse protein mixture by tandem mass spectrometry: 1) gel-free approach, referred as shotgun proteomics [73, 105, 106], in which purified proteins are directly digested in solution, the resulting tryptic peptides are separated by one-dimensional or multidimensional chromatography and on-line injected into a tandem mass spectrometer via nano-ESI (this method is also known as nanoflow liquid-chromatography-tandem mass spectrometry (nanoLC-MS/MS)) and 2) gel-based approach, referred as Gel-LCMS, in which proteins are first separated by one or two-dimensional electrophoresis, enzymatically digested in-gel with proteolytic enzymes and the extracted peptides are either directly analyzed by tandem mass spectrometry or subjected to one or multidimensional chromatographic separation prior to mass spectrometric analysis [107].

1.3.2.1 Gel-free approach

Given the limitations of two-dimensional gel electrophoresis, alternative methodologies employing multidimensional chromatography for the separation of complex peptide mixtures prior to analysis by MS have found preferential application in many proteomic studies. Multidimensional separation couples two or more different separation methods. Greater chromatographic resolution obtained by multidimensional separation methods can be achieved by taking into consideration criteria established by Giddings et al. [108-110], who demonstrated that the overall peak capacity of multidimensional separations is the product of the peak capacities in each independent dimension only if the separation dimension are orthogonal and component separated in one dimension remain separated in any additional separation dimension [105].

The multidimensional peptide separation methods reported following Giddings' criteria include chromatographic techniques based on hydrophobicity, charge, molecular weight, or functionality of peptides. For instance, the separation of peptide mixtures by 2 D LC/LC methods can be performed using several orthogonal combinations such as strong cation exchange / reversed phase liquid chromatography (SCX/RPLC), anion exchange chromatography / reversed phase liquid chromatography (AE/RPLC), and affinity chromatography / reversed phase liquid chromatography (AFC/RPLC). Typically, the second dimension is performed by RPLC because the mobile phase is compatible with the mass spectrometric analysis [71].

The most prominent and commonly used strategy, however, applies SCX (separation on the basis of charge) coupled to RPLC (separation on the basis of hydrophobicity). There are two main approaches, offline and online. In offline separation, developed by Link et al. [73], the first dimension (SCX) is not directly coupled to the second dimension (RP) or SCX-RP. Fractions from the SCX column are collected and later subjected to the RP column. The online approach, refined by Washburn et al. [106] employs coupling the two chromatographic methods together so that the eluent from the first dimension (SCX) is directly eluted onto the second dimension (RP) or SCX/RP, thus avoiding the need for fraction collection. Online approaches are substantially faster than off-line approaches, and sample loss is minimized due to the direct coupling of the two dimensions. There are different variations of the online approach such as using separated columns for the SCX and RP connected by switching valves, or using multidimensional protein identification technology (referred as MudPIT), where the SCX and RP stationary phases are packed together in the same microcapillary column [105, 106]. To enable desalting of biological samples, which typically contain urea and other salts for optimal protein digestion, triphasic and split-three-phase [111, 112] column were designed. These developments enabled direct loading of samples on the column without offline desalting, which leads to sample loss and longer analysis times.

MudPIT technology has become an important technique in bottom-up proteomics. It has been applied in a wide range of application, ranging from extensive proteomic analysis of different organisms or their subcellular components [105, 106, 113-115] to characterization of multiprotein complexes [116-118] and their quantification [119-122].

High sensitivity in shotgun analysis is achieved by microcapillary column (50-100 µm i.d. columns, operating at 100-350 nl/min) [123], which were first introduced by Hunt et al. [124]. Further advances in nanoHPLC technology will also bring great improvements in this field. Giddings demonstrated the importance of orthogonality and how this increases the number of theoretical plates in a given analysis [108-110]. Another way to increase the number of theoretical plates in a chromatography analysis is to apply smaller particle size, which then requires higher pressures for chromatographic analysis. Ultrahigh-pressure reversed phase liquid chromatography (UHPLC) has become an active area of research [125-128]. Smaller reversed phase particles have been synthesized and applied in UHPLC [128, 129] and early efforts to implement an ultrahigh-pressure MudPIT system have been promising [130]. However, to have a fully integrated orthogonal two-dimensional UHPLC shotgun proteomics system, research into small particle and high-pressure resistant strong cation exchange particles is required [71].

1.3.2.2 Gel-based approach

Gel-LCMS is a powerful method in the analysis of complex protein mixtures [131-135]. From practical point of view gel-based approach has several advantages compared to the gel-free shotgun strategy. First, gel electrophoresis separates proteins into narrow mass range bands, significantly increasing the dynamic range of proteomic analysis. Second, in gel-based approach detergents and buffer salts, which are not compatible with mass spectrometry (especially based on ESI), are washed out from the gel matrix, making this method appropriate to high-throughput MALDI-MS and nanoES MS/MS analysis of isolated protein bands. And finally, gels can be stored for years without noticeable changes in pattern of tryptic peptides and in their recovery [136].

Although in-gel digestion is well established for bottom-up proteomics and has been routinely used for more than a decade [137], it has significant limitations. Its major limitation is poor peptide yield that limits the analysis sensitivity. One of the basic factors responsible for reduced peptide recovery is the limited efficiency of in gel-digestion due to slow diffusion of trypsin molecules inside the gel matrix [2]. Therefore, to achieve better efficiency much higher concentrations of enzymes (compared to in-solution proteolysis) have to be applied, resulting in significant autolytic background.

Next, each step of gel processing, such as performing of electrophoresis, gel staining, cutting of protein bands, in-gel digestion and extracting of tryptic peptides increase the risk of contaminating samples with keratins or other contaminants, so enhancing chemical noise in analyzed samples.

And, finally, compared to shotgun approach, which is based on in-solution digestion of proteins and enables protein identification in relatively short time, in-gel digestion is a time-consuming procedure, which requires overnight protein cleavage and additionally pre-digestion sample preparation.

Significant improvement of in-gel digestion was achieved by Havliš et al. [2]. He addressed mentioned limitations using porcine trypsin with methylated ε-amino group of lysine residues, which represent better thermostability and autolysis resistance (by kept cleavage specifity) than its unmodified form and commonly used in proteomics bovine trypsin. Havliš et al. developed accelerated in-gel digestion protocol, which considerably reduced the proteolysis time (down to 0.5-1h instead of overnight incubation) and improved the recovery of digestion products. Thus, it has become apparent that further development of autolysis resistant and thermostable trypsin conjugates (provided that they maintain their catalytic activity and cleavage specifity) would enable major advance towards fast and flexible protein analysis.

Trypsin undergoes rapid autolytic inactivation at basic pH (corresponding to pH optimum of the enzymatic reaction) presumably due to hydrolysis of C-terminal lysine and arginine peptide bonds. Such autolysis may be prevented or minimized by chemical modification of the ε-amino group of lysine residues and guanidino group of arginine residues. A number of experiments to modify these amino acid residues were carried out in order to stabilize trypsin [138-141]. It was shown that reductive methylation increases autolysis resistance and thermostability [142] of trypsin, without strong impact on its catalytic activity and without altering of its substrate specificity. However, in many cases chemical modification of enzymes has been reported to provoke significant losses of catalytic activity [138, 140].

The interest for modifying enzymes with sugar moieties has been raised because of the better stability and functional properties showed by the naturally occurring glycoenzymes [143]. Their stability against thermal inactivation is assumed to derive mainly from the hydrophilization of the non-polar areas of the enzyme, as a result of the covalent attachment of the oligosaccharide to exposed lysine residues [140, 144]. Hydrophilization hinders thermal denaturation associated with the formation of new intra- and inter-molecular hydrophobic interactions in the course of thermal treatment [144].

In order to further improve the thermostability of trypsin in line with fast digestion approach developed by Havlis et al., Šebela et al. synthesized trypsin conjugates by coupling oligosaccharides to its lysine residues and characterized them bioanalytically [1]. Trypsin conjugates significantly increased thermostability and autolysis resistance of trypsin, without affecting its cleavage specifity, revealing their great potential for accelerated digestion of proteins both in-solution and in-gel.

1.4 Analysis and validation of proteomic data produced by nanoLC-MS/MS

Nanoflow liquid chromatography-tandem mass spectrometry (nanoLC-MS/MS) is an automated, high-throughput analytical method and generates thousands of tandem ion spectra in a single analysis [105, 112]. The correct assignment of these MS/MS spectra to peptide sequences and identification of analyzed proteins is a complex process, which involves pre-processing of raw data, peptide/protein identification and validation of the obtained results [145-147].

1.4.1 Pre-processing of raw data

Successful protein identification depends on good pre-processing of mass spectrometric data. The main goal of pre-processing of MS/MS spectra is to increase the specifity, sensitivity and accuracy of automatic database searches. Pre-processing includes peak detection, noise reduction, and monoisotopic peak determination. These parameters strongly depend on the quality of the acquired data [148].

Analysis of complex peptide mixtures by shotgun approach results in huge number of fragment ion spectra, while many of them are redundant [149], because of repeated fragmentation of highly abundant peptides. This dramatically increases the complexity of data-analysis, in terms of computational processing time required and time required for validation of the obtained results. To overcome this problem significant efforts have been undertaken to develop algorithms, which enable clustering and merging of redundant tandem mass spectra [149-153]. Further Zhang et al. introduced software capable to recognize spectra generated from cofragmentation of two or more peptides [154].

Another challenge in analysis of acquired spectra derives from presence of background peaks, which complicate database searches. Low-energy CID fragmentation generates predominantly *a*, *b*, *y* ions and their derivates, which have lost ammonia (-17 Da, a*, b* and y*) and water (-18 Da, a°, b° and y°). However, MS/MS spectra contain many more peaks. These can result not only from isotope variants and multiply charged replicates of the peptide fragmentation products but also from unknown fragmentation pathways, sample-specific or systematic chemical contaminations or from noise generated by the electronic detection system [155, 156]. The presence of this background not only complicates spectrum interpretation by increasing computational time, but also might lead to incorrect protein identification. To address this problem considerable efforts have been made to study in depth peptide fragmentation chemistry [157-159] and to develop algorithms for detection and transformation of multiply charged peaks into monoisotopic peaks, removal of heavy isotope replicates, and random noise [156, 160]. Sophisticated charge determination software [161, 162] were introduced to this end. Since acquired data contain high number of poor quality spectra, which are often of non-peptidic nature, several strategies have been addressed to measure the quality [163] of tandem mass spectra filtering low quality spectra prior database searching [164-168].

Analyzed samples also contain peptides from common contaminants like human and sheep keratins, proteolytic enzymes, antibodies, GST etc. Many of these sequences are either not present in a database or scattered through a large number of partially redundant database entries. When abundant, they also give rise to a large pool of polymorphic sequences, orifice fragmentation products, sodium adducts etc. [169]. Therefore, it would be advantageous to remove these spectra prior to database searches. To this end computational algorithms have been developed, which recognize and remove these background spectra [152, 169], so decreasing the amount of data and avoiding possible false positives.

1.4.2 Peptide/protein identification based on database searching

A large number of computational methods have been developed to assign peptide sequences to acquired tandem mass spectra. Generally, there are three ways to identify analyzed proteins: 1) database searching involves assignment of acquired spectra to theoretical spectra *in silico* predicted for each peptide contained in a protein sequence database [170-176]; 2) *de novo* sequencing derives peptide sequences directly from MS/MS spectra based on peptide fragmentation rules [159, 177-182]; 3) hybrid approach, pioneered by Mann [183] involves *de novo* identification of short sequence tags followed by ‘error-tolerant’ database searching [184, 185].

Database searching is most suitable approach for large-scale proteomics and several database search programs have been developed to this end, such as MASCOT [170], SEQUEST [172], X!TANDEM [171], ProbID [173], Phenyx [174] etc. All these algorithms rely on comparison of the acquired MS/MS spectra with theoretical spectra predicted from a sequence database using common peptide fragmentation rules. A number of search parameters need to be considered here, for instance, searching database, proteolytic enzyme specifity, amino acid modifications (stoichiometric, called "fixed" or non- stoichiometric, called "variable"), and mass tolerance of precursor and fragment ions. Database searching programs apply scores, which represent the degree of similarity between the acquired and the theoretical spectrum, and therefore serve as the primary discriminating parameter for separating correct from incorrect identifications [186].

Automated database search enables fast large-scale protein identification. However, high number of acquired MS/MS spectra remains unmatched or matches peptides with low scores, resulting in proteins that were not actually in the sample – false positives and leaving out proteins that were in the sample – false negatives. There are several reasons for this problem [187].

First, database searching approach only enables identification of those peptides that are present in the searched sequence database. Since the peptide molecular weight is used as a filter to derive candidate sequences from a database, an incorrect molecular weight will provide incorrect sequences. Thus, in *ex vivo* / *in vitro* modified peptides (with oxidized methionines or carbomidomethylated cysteines residues), which were not specified by database search or peptides derived from post-translationally modified proteins remain unassigned or might match incorrect peptides. The same problem concerns protein identification from organisms not well represented in any sequence database. Even, for organisms with completely sequenced genomes, sequence polymorphisms can still cause difficulties, since these are sometimes indicated only as annotations, rather as separate sequence entries. Common background proteins might also lead to false positives by database searching in small species-restricted databases.

Second, typically database search is performed under specification of the applied proteolytic enzyme. Peptides derived from cleavage by another proteolytic activity present in the sample or from fragmentation of intact peptide ions in the ion source prior to mass analysis will not be correctly identified.

Third, since the peptide mass is a filter parameter by database searching, an incorrectly determined peptide mass (for example, incorrectly called monoisotopic peak) or charge state of a peptide ion selected for fragmentation will provide incorrect sequence candidates, leading to possible false positive identifications or unassigned spectra.

Fourth, database search is based on a simplified representation of the peptide ion fragmentation rules. Unexpected fragmentation pathways complicate peptide identification [188].

Fifth, for better sensitivity QTOF and IT instruments are typically operated with an isolation window for precursor ions of 3-4 Da. Therefore, it is uncommon that acquired MS/MS spectra contain fragment ions from coeluted precursor ions that are close in mass.

Sixth, dirty solvents used in HPLC might contain alkali metal cations, which can build sodiated peptide ions. This lead not only to increased peptide mass, but also changes the fragmentation pattern compared to unmodified peptide ions.

Seventh, high number of acquired spectra derives from non-peptide contaminations, resulting in incorrect peptide identifications or unassigned spectra.

Generally, the accuracy of peptide/protein identification strongly depends on the performance of the applied mass spectrometer, data quality, and the appropriate chosen database [186].

1.4.3 Protein identifications with borderline statistical confidence

A major limitation in identifying peptides from complex mixtures by shotgun proteomics is the ability of search program to accurately assign peptide sequences to the acquired MS/MS spectra. This problem is addressed by all search engines by applying sophisticated scoring techniques, which evaluate the probability of false positive identifications. MASCOT algorithm, for instance, applies probability based scoring and typically establishes threshold scores which reflect 95% (usually used) confidence level when searching data from peptide mass fingerprints and tandem mass spectra. However, in large scale proteomics projects, when thousands of peptides are identified, this level of confidence may be unsatisfactory, resulting in false positive peptide/protein identifications. On the other hand, these threshold scores depend on the database size. To confidently identify a protein in a comprehensive database (even at this moderate confidence level), the ion scores of analyzed peptides should exceed a relatively high threshold score. Given the complexity of analyzed mixtures (up to 10 orders of magnitude [70]) and limited sensitivity of current analytical instruments, many of acquired MS/MS spectra are of not sufficient quality. Thus, identification of high number of proteins relies on matching one or two marginal quality mass spectra with scores far below this threshold score (for comprehensive database). Rejecting of these hits would significantly increase number of false negatives. On the contrary, for database searches in small species-restricted databases threshold scores are lower. Accepting hits corresponding to these threshold scores would inevitable result in increased number of false positives.

The problem of borderline hits is more pronounced when peptide sequencing is performed at low femtomole level, where peptide precursors are often contaminated by co-selected background ions, and are affected by poor ion statistics [152]. Moreover, background proteins originated from exogenous species, such as human and sheep keratins, fragments of proteolytic enzymes, antibodies, fragments of expression vectors or protein from host organisms contribute to the false positive rate if database searching is performed against a small species-restricted sequence database. Independent of the applied algorithm [170-174], database searching is a probabilistic process, in which the confidence of hits is evaluated by the comparison of some matching quality scores against empirical or semiempirical statistical significance thresholds.

Manual evaluation can not be regarded as appropriate validating tool in such big scale analyses. Thus, revealing real hits (false negatives) among ocean of ambiguous protein identification is a challenging task in today’s proteomics, requiring improvement statistical methods of database searching engines, and deeper understanding of peptide fragmentation pathways and their impact on the accuracy of spectrum-to-sequence matching [158, 159, 189-191]. It is also important to independently validate borderline hits irrespectively of statistical properties of both the spectra dataset and sequence database.

Several strategies have been developed to validate ambiguous hits based on additional information, such as agreement between sequence composition of the identified peptide and its chromatographic behaviour [192, 193], probability of missed cleavages [194] or exact mass measurements [192]. In addition, some methods have applied intensity information in validating of data [191, 195, 196].

However, in many proteomics studies manual validation of borderline identifications is still regarded as the method of choice. The main weakness of this approach is that it is completely based on subjective decision of the analyst. It is therefore not surprising that high number of proteomics publications represent ambiguous protein hits, whose identifications are often based on matching a single peptide, of completely non-tryptic termini [197, 198]. Recently introduced "Manual Analysis Emulator" (MAE) was developed to automate key aspects of manual analysis, minimize subjective decisions, and enable high-throughput processing [199]. The method is based on the *de novo* sequencing program MassAnalyzer, developed by Zhang et al. [159, 200], which simulates MS/MS spectra including relative fragment ion intensities. This program applies new kinetic model for peptide fragmentation integrated from recently investigated gas phase chemistry mechanisms of peptides [30, 201, 202].

To address the issue of database independent validation, Savitski et al. [168] introduced a new scoring method (S-score), which utilizes the advantage of combined use of complimentary fragmentation techniques collisionally activated dissociation (CAD) and electron capture dissociation (ECD). S-score is based on the maximum length of the peptide sequence tag predicted from CAD and ECD data, enabling confirmation of some of the below threshold hits, and revealing false positives and modified sequences. The quality of MS/MS spectra assessed by S-score also allows poor data to be filtered out before the database search, speeding up the data analysis and eliminating a major source of false positive identifications.

Despite undertaken efforts to develop methods [168, 199], which provide database independent validation of hits with borderline statistical confidence, there is a particular need in validation algorithms applicable in high-throughput proteomics.

1.4.4 Statistical assessment of peptide assignments in large-scale datasets

One of the first strategies to separate correct from incorrect peptide assignments in data analysis was application of *ad hoc* filtering criteria based upon database search scores and some properties of the assigned peptides often in conjunction with manual validation [193, 203-206]. However, the numbers of rejected correct identifications and accepted false identifications that result from applying such filters are not known. Moreover, the obtained score distributions depend on several factors, such as the performance of the mass spectrometer, data quality, and the size of the database. Therefore, application of the same thresholds to data from different experiments would result in different (and unknown) error rates, making comparison between datasets practically impossible [186]. Thus, consistent and reliable interpretation of data to enable the comparison of results from different experimental groups requires robust statistical methods to validate peptide assignments to MS/MS spectra [188].

Several statistical methods for validating of peptide identifications have been developed on top of existing database search tools [207-211]. Generally, the global statistic approaches can be broadly grouped into two categories: target-decoy searching and empirical Bayes approaches [186].

The first strategy relies on searching target-decoy databases, and computes an optimized cut-off score for each database. Two different types of searches have been described: in the first step the MS/MS spectra are searched against the database of interest and a randomized database independently [212]. In the second step the original database and a randomized database are joined (concatenated) and searched simultaneously [211]. Peptide assignments are then filtered using various cut-offs, and the corresponding FDR for each cut-off is estimated as $2N_D/N$, where N is the number of peptide matches with scores above the cut-off and N_D is the number of matches to decoy sequences among them. Target-decoy approach assumes that matches to decoy peptide sequences and false matches follow the same distribution and has been proposed to be very robust method. It is simple and can be applied in large-scale proteomics by evaluation of data generated by LC-MS/MS analyses. However, doubled database search time should be considered. Still, the serious concern is whether reversing or randomizing sequences can provide an accurate assessment of the distribution of false peptide matches when many of those are known to be sequences homologous to the true peptides rather than completely random sequences [186].

The second statistic approach exemplified by PeptideProphet algorithm developed by Keller et al. [207] is based on linear discrimination analysis and estimate the accuracy of peptide assignments to tandem mass (MS/MS) spectra made by database searches [188]. In this approach each peptide assignment to a spectrum is evaluated with respect to all other assignments in the dataset, including necessarily some incorrect assignments. It uses the observed information about each assigned peptide in the dataset, learns to distinguish correct from incorrect assignments and, finally, computes the probability for each assignment. By evaluation of peptide assignments PeptidePhrophet typically includes database search scores, the difference between measured and theoretical peptide mass, the number of termini consistent with the type of enzymatic cleavage used, and the number of missed cleavage sites. In addition, PeptideProphet can apply auxiliary features, such as peptide retention time [212, 213], pI of identified peptides [75, 214], presence of special amino acids (for example, cysteine in the case of avidin affinity purification of peptides containing biotinylated cysteins), expected number of missed cleavages (for example, missed cleavages can occur in the presence of acidic groups near cleavage site, otherwise they can result due to cleavage sites being adjacent to one another), providing valuable information for validation of ambiguous hits [188].

The described statistical approaches are widely used in proteomic data analysis. However, they both evaluate thresholds of statistical significance and do not imply the validity of individual spectrum-to-sequence matches. These methods do not replace the need for data base independent validation strategies.

1.4.5 Validation of protein identification: protein interference problem

A separate problem in data analysis is validation of protein identifications. In bottom-up proteomics proteins are digested prior to LC-MS/MS and their identification is based on analyzed peptides. The connectivity between peptides and proteins is usually quite straightforward when analysed protein mixtures are not complex and separated by 2D electrophoresis (additionally information of protein mass and its isoelectric point is available). In case of complex protein samples analyzed by MuDPIT technology this connectivity is lost, interfering protein identities (when one particular peptide can be assigned to multiple proteins) from the set of identified peptides becomes a serious problem. This problem arises from protein paralogues, splicing variants, or redundant entries within the database. Here again, statistical methods were developed [208, 210, 215]. The statistical method of Nesvizhskii et al. used in software tool ProteinProphet computes probabilities that a protein is present in the sample by combining the probabilities that corresponding peptides are correct [215]. Here individual peptide probabilities are aligned for observed protein grouping information.

1.4.6 *De novo* sequencing and homology searching

An alternative approach to peptide/protein identification is *de novo* sequencing, where peptide sequences are directly derived from fragmentation spectra without recourse to a sequence database. Significant efforts have been invested into development of *de novo* sequencing algorithms [159, 177-182]. However, *de novo* sequencing is difficult and error-prone approach that typically produces ambiguous results [187]. There are several reasons for it.

First, there are difficulties in differentiating between some amino acids of identical (leucine and isoleucine) or nearly identical masses (e.g. glutamine/lysine and phenylalanine/oxidized methionine, which, however, can be resolved by instruments with high resolution and mass accuracy). Moreover, some pairs of amino acids have identical or nearly identical masses to certain amino acid residues. Second, ion series are rarely complete, since fragmentation does not occur at every peptide bond. In addition, fragment ions are present in varying abundances (often below noise level), in many cases with associated losses of water and/or ammonia, what complicates *de novo* sequencing. Third, it is usually not known whether an ion contains the C- or N-terminus of the peptide. To address this problem Shevchenko et al. [216] introduced isotopic labeling of C-terminus by trypsin proteolysis in 50 % $H_2{}^{18}O$, which labels y ions as doublets separated by 2 Da, so helping to connect the observed ion series to the correct termini.

It seems to be reasonable that *de novo* sequencing can be combined with homology-based searches, providing complementary validation approach to database searching [217]. However, *de novo* interpretation of tandem mass spectra results in many relatively short (usually 6-12 amino acid residues) sequence proposals, which are highly redundant and error-prone. Conventional database search algorithms such as BLAST [218] or FASTA [219] are optimized for accurate and long (>35 amino acid residues) sequence queries, where amino acid permutations (such as leucine/isoleucine and glutamine/lysine), gaps or insertions/deletions are strongly penalized. In addition, homology searching is computationally demanding. To address these difficulties existing sequence alignment algorithms have been modified in order to match *de novo* sequences to protein sequence databases. For example MS-BLAST [220], MS-Shotgun [221], CIDentify [222] and FASTS [223] can be used to align *de novo* sequences to database homologues using highly efficient sequence alignment algorithms.

The idea to use *de novo* sequencing for validation of ambiguous results is not new. Taylor et al. [217] applied automated program Lutefisk in conjunction with a homology-based database search program CIDentify, which uses a modified FASTA sequence comparison algorithm to screen the sequences produced by automated interpretation of low-energy CID spectra, in validation of database searches. The authors also showed that this strategy can be used for identification of homologous protein families from data obtained from unknown proteins [222] as well as by characterization of posttranslational or chemical modifications and peptide originated from nonconsensus proteolytic cleavages. However, because of rapid growth of sequence databases, the throughput of the approach is limited by the relatively long running times required by the modified FASTA algorithms [221-223]. In addition, the significance of hits in these algorithms depends not only on the number of matched peptides, but decreases with the increasing number of redundant peptide sequence candidate in the query.

Mass spectrometry driven BLAST (MS BLAST), developed by Shevchenko et al. [220, 224], utilizes degenerate, redundant and partially inaccurate peptide sequence candidates obtained by *de novo* interpretation of tandem mass spectra. MS BLAST is web accessible program, which is operated at servers of very high computational capacity and can be applied for high-throughput analysis of data. MS BLAST doesn't employ original statistical evaluation procedure of classical BLAST (no E-values or p-values) instead it uses a scoring matrix optimized for peptide sequences produced by *de novo* sequencing of MS/MS spectra [225].

1.5 Quantitative mass spectrometry in proteomics

Mass spectrometry is increasingly used for quantitative proteomic profiling of complex biological samples. Quantitative proteomics is important to provide fundamental understanding of biological processes because the kinetics/dynamics of the cellular proteome is described in terms of changes in the concentrations of proteins in particular compartments [226]. Generally, the quantification strategies can be divided into two categories: 1) quantification using stable isotope labeling, including metabolical, enzymatical labeling, labeling by chemical means or provided by spiked synthetic peptide standards and 2) label-free quantification using spectral counting or spectral feature analysis (Figure 1.2) [227].

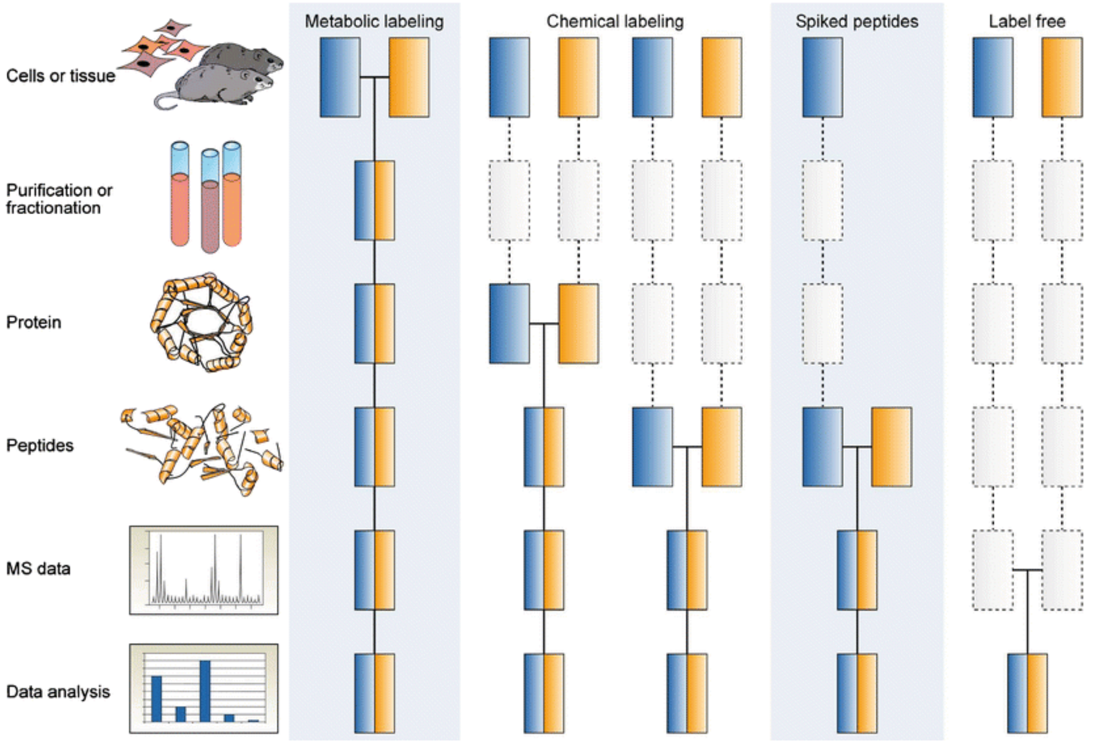


Figure 1.2. Common quantitative mass spectrometry workflows.

Boxes in blue and yellow represent two experimental conditions. Horizontal lines indicate when samples are combined. Dashed lines indicate points at which experimental variation and thus quantification errors can occur (adapted from [227]).

1.5.1 Stable isotope labeling

One commonly used approach in bottom-up proteomics employs stable isotope labeling (^{12}C vs. ^{13}C, ^{14}N vs. ^{15}N, ^{2}H vs.^{1}H), allowing comparison of peptides between samples. Stable isotopes labeled peptides are chemically identical to their native counterparts and therefore have similar behaviour during chromatographic and mass spectrometric analysis. Isotope labels can be introduced into amino acids 1) metabolically, 2) chemically, 3) enzymatically or, alternatively, by spiking of synthetic peptides.

Metabolic labeling involves *in vivo* incorporation of stable isotopes into the proteins in special media containing these isotopes. In this method cells are grown in two different media containing ^{14}N or ^{15}N isotopes, then combined and analyzed by MS [228-230]. The main disadvantage of this strategy is not predictable mass shift, since the method labels all nitrogen atoms of the backbone and side-chains. In an alternative approach, termed “stable isotope labeling with amino acids in cell culture (SILAC)” [231], proteins are labeled *in vivo* by growing cells in media containing isotopically labeled amino acids, such as ^{2}H-leucine, ^{13}C-lysine, ^{13}C-tyrosine, ^{13}C-arginine[232-234]. This method has become popular because of the predictability of the mass shift. Generally, isotope labeling *in vivo* has the advantage that it happens in the early stage of preparation, so reducing variance between samples. A disadvantage, however, is that this approach can not be applied for analysing biological samples that cannot be grown in culture, such as tissues and body fluids [235].

In vitro labeling approach involves incorporation of stable isotopes by chemical reaction at the amino- or carboxyl- terminal of targeted peptides, or on specific amino-acid residues, such as cysteine, lysine, tyrosine etc. ‘Isotope-coded affinity tags’ (ICAT) approach introduced by Gygi et al. [236] applies a reagent consisted of biotin affinity tag for selective purification, a linker that incorporated stable isotopes (^{1}H or ^{2}H) and an reactive iodoacetamide group, which reacts with cysteinyl thiols. This method has been significantly improved by introducing an acid-cleavable linker that allows removal of the large affinity tag prior to MS and incorporation of carbon-13 instead of deuterium that prevents possible chromatography shifts [237-240]. ICAT quantification strategy was applied to variety of species [204, 241]. However, ICAT is not suitable for quantifying of proteins, which do not contain any cysteine residues. Thus, many biologically ‘interesting’ protein changes might remain uncharacterized by this approach [242].

Other groups of reagents targetN-terminus of peptides and amino group of lysine via the very specific N-hydroxysuccinimide (NHS) chemistry or other active esters and acid anhydrides [243-247], as well as via methylation of lysine residues by formaldehyde via Schiff base formation and subsequent reduction by cyanoborohydrate [248-250], iTRAQ reagent (isotope tags for relative and absolute quantification) has to be pointed out [251]. In contrast to ICAT and similar mass-difference labeling strategies, quantitation is performed at the MS/MS stage rather than in MS. iTRAQ reagent consists of a reporter group, a balancer group and a peptide reactive group, which reacts with primary amino groups of peptides. The specifity of this approach is that the mass of balancer and reporter group remains constant, whereas the reporter group ranges from 114 to 117 Da, and balancer group ranges from 28 to 31 Da, making differently labeled peptides isobaric (they have similar chromatographic behaviour) (Figure 1.3). During CID, the reporter group ions fragment from the backbone peptides, representing different masses from 114 to 117 Da, allowing multiplexed quantification.

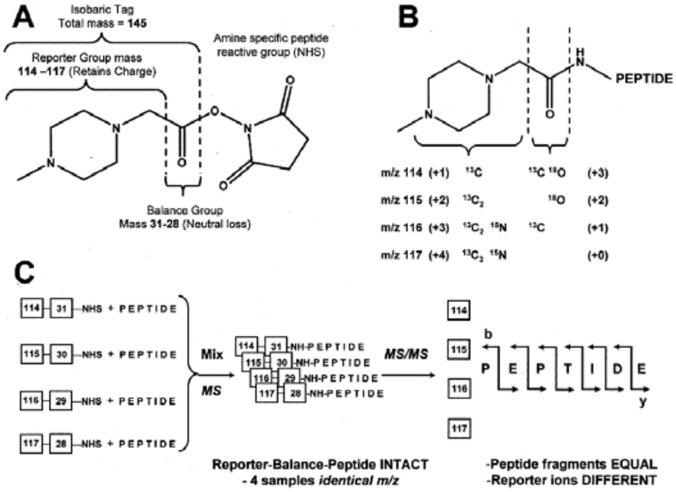


Figure 1.3. Strategy for protein quantification by iTRAQ.

(A), structure of iTRAQ reagents. (B), differentially labeled reporter and balancer groups. (C), four isobaric combinations with four different reporter group masses. Following CID, the four reporter group ions appear as distinct masses (114–117 Da); adapted from [251].

Carboxylic acids in side chains of glutamic and aspartic acid residues as well as the C-termini of peptide chains can be isotopically labeled by esterification used deuterated alcohols [252]. This reaction has become attractive especially for quantification of phosphopeptides, because esterification improves the specifity of their enrichment procedure [253]. Several stable isotope labeling methods have been developed for quantification of phosphorylated [254-257] and glycosylated peptides [258].

Stable isotopes can be introduced enzymatically to the C-termini by proteolytic digestion of proteins in $H_2{}^{18}O$, or after proteolysis by incubation of the obtained peptides with a protease in $H_2{}^{18}O$, resulting in mass shift of 2 Da per ^{18}O atom [259-261]. Acid- or base-catalyzed back-exchange can occur at extreme pH values, but under mild acidic conditions ^{18}O-containing carboxyl groups of peptides are stable [262]. The main disadvantage of this method is that the full labeling is seldom achieved and that incorporation of one or two oxygen atoms depends on the nature of peptide, complicating the analysis [263, 264].

A method known as AQUA applies isotope-labeled synthetic standards [265]. Known quantities of labeled peptides added to protein digests provide information for absolute quantification. Application areas of this approach are analysis and validation of potential biomarkers in a large number of clinical samples [266] and determination of protein stoichiometry in protein complexes [267, 268]. However, this approach can not be used in large-scale quantifications, because of the high manufacturing cost of standard peptides, which have to be chemically synthesized in stable isotope−labeled form and independently quantified. This approach has been refined by constructing synthetic genes that express artificial proteins what are concatemers of tryptic peptides for several proteins or group of proteins [269].

One practical limitation of the AQUA approach is that by given complexity (first and foremost high dynamic range) of analyzed tryptic digests it is difficult to decide how much of the labeled standard should be added to a sample; this amount might significantly vary for all proteins of interest. Another limitation is the specifity of the spiked standard as there are likely multiple isobaric peptides present in the mixture. Both of these issues can be improved by multiple reaction monitoring (MRM) in which the mass spectrometer monitors both the intact peptide mass and one or more specific fragment ions of that peptide [270]. Application of auxiliary information, such as retention time, peptide mass eliminates ambiguities in peptide assignments and extends the quantification range to 4-5 orders of magnitude [271].

Although protein quantification using stable isotopic labeling has been proved as accurate, sensitive and reproducible method it has several limitations. As first, labeling with stable isotopes is often very expensive, and some labeling procedures involve complex sample preparations. Second, labeling methods make acquired LC-MS spectra more complex due to the presence of additional isotopic peaks, which often overlap with co-eluting components of similar masses, complicating peak detection and quantification. And finally, chemical labeling approaches are prone to side reactions (e.g. thiol reactions of serine and threonine residues with iTRAQ reagent), leading to unexpected products [227].

1.5.2 Label-free quantification

Label-free quantitation strategies are promising alternatives to stable isotope labeling approaches. Their advantages are simple and less expensive sample preparation, lower sample complexity, applicability to any samples, including tissues and ability to quantify and compare multiple samples. There are two fundamentally different strategies for label-free quantification: the first one measure and compares mass spectrometric signal intensities of peptide precursor ions of a given protein [272-276] and the second one counts and compares the total number of MS/MS spectra of any peptide for a given protein [106, 226, 277, 278].

Spectral feature analysis is a quantification approach, which is based on measuring and comparing the mass spectrometric signal intensities of peptide precursor ions of a particular protein. This is typically done by creating extracted ion chromatograms (XICs) for the mass to charge ratios determined for each peptide. The intensities for each peptide in a given sample can be compared with intensities of the corresponding peptides in other samples, enabling relative quantification of multiple samples. Integrated peak areas, however, can be influenced by different factors including ion suppression, limited ion trapping capacity of mass spectrometers, or simply by the parameters applied to create extracted ion chromatograms, *e.g. m/z* tolerance, background subtraction etc [279]. Spectral feature analysis is not applicable to low abundant proteins, due to difficulties to accurately define peaks and signal to noise ratio [280]. This quantification approach requires stringent statistical methods and replicate analyses and strongly depends on the accuracy of the mass measurement and the chromatographic reproducibility [281]. It is therefore advantageous to apply high mass accuracy instruments, which minimize the interference of peptides with close masses. The chromatography should be also optimized for better resolving of peptides, especially in complex protein mixtures. Special software have been developed to accurately align and profile features between many LC-runs [275, 282-285]. In addition, the right balance between acquisition of survey and fragment spectra has to be found, since better quantification accuracy, which requires multiple sampling of the chromatographic peaks by survey MS, means poorer proteome coverage (high proteome coverage can be achieved by extensive peptide sequencing by tandem mass spectrometry). To address this issue the analysis can be performed in two steps: in first experiment the instrument is adjusted to identify as many peptides as possible and in the second experiment mass spectrometer operates only in MS-mode to optimize sampling of peptide signals. An another approach refers to differential feature detection: here as first a survey scan is performed to profile ions showing differences and subsequently the sample is reanalyzed by tandem MS to identify those ions [276, 286].

The ability to determine the absolute concentration of a protein (or proteins) in protein complexes is important to understand their stoichiometry. The absolute amount of a protein can be obtained using synthetic labeled internal standards chemically identical to the proteotypic peptides generated by protein proteolysis [266, 269]. Due to the limitations of this strategy, mentioned before, there is a particular need in development of label-free quantification methods to estimate absolute quantities of proteins. Recently Silva et al. [287] showed that the average MS signal response for the three most intense tryptic peptides per mole of protein is constant. Given an internal standard, this relationship is used to calculate a universal signal response factor, so providing method for absolute quantification.

Spectral counting approach is based on the observation that the number of acquired MS/MS spectra for sequenced peptides depends on the quantity of a given protein. This method sums the total number of tandem mass spectra of any peptide of a given protein observed at different charge states, or in different chromatographic fractions. The protein abundance is then estimated from the number of obtained MS/MS spectra for a corresponding protein normalized to its length or expected number of tryptic peptides [226, 279]. Spectral counting enables relative quantification by comparing the protein abundance between different experiment sets. In contrast to quantification by peptide ion intensities, spectral counting benefits from extensive MS/MS data acquisition across LC-MS/MS experiment. Dynamic exclusion is a commonly used tool in tandem mass spectrometry, which employs exclusion of ions that have already been selected for fragmentation, enabling fast collection of information without repetitions. However, it is disadvantageous for accurate protein quantification by spectral counting. Spectral counting approach is still controversial, mainly because it assumes linear response for different proteins [227]. In fact, the response is varying for different peptides due to their distinct physical properties. Reasonable results can be obtained when sufficient number of MS/MS spectra was obtained for a given protein. Old et al. showed that protein ratios 2-fold or greater could be estimated, however, to achieve high confidence at this level ≥4 spectra/protein were required [281]. On the other side, saturation effects can be obtained at higher spectral counts, complicating quantification of complex protein mixtures with high dynamic range. Nevertheless, the practical utility of spectral counting approach has been demonstrated in several applications [279, 288].

It should be noted, when comparing both label-free quantification methods, that spectral counts strategy more accurately quantify large changes in abundance, whereas spectral feature approach provides better estimates of smaller changes [121, 281]. Although, both label-free quantification methods are not as precise as stable isotope labeling, they can be used to address many biological questions, including those cases where labeling is not possible. In addition, label-free methods provide higher dynamic range of quantification than stable isotope labeling strategies, since the complexity of a sample significantly increases by adding of labeled internal standards [227].

1.6 Questions and aims of the thesis

LC-MS/MS analysis often in combination with 1 or 2 D gel electrophoresis has been the standard method for identification and quantification of proteins in bottom-up proteomics. Despite continuous improvements of MS instrumentation and software, several bottlenecks have been recognized, such as:

1) low efficiency of in-gel digestion, which requires long processing times and results in poor peptide yield, strongly contaminated with autolysis products of the used protease,

2) large number of protein identifications with borderline statistical confidence at the edge of sensitivity and finite dynamic range of MS instruments.

To address these issues I set the following goals for my thesis work:

1. Evaluate the performance of trypsin derivates modified with oligosaccharides, which offer better thermostability [1] than unmodified commercially available bovine trypsin.

2. Study the kinetics of in-gel digestion of proteins by glycosylated trypsins, in order to evaluate how the reaction temperature, enzyme concentration and digestion time affect the yield of digestion products [2].

3. Establish a reliable, automated and database independent method for rapid validation of protein identifications with borderline statistical confidence and test its performance in large-scale protein identifications.

2 RESULTS AND DISCUSSION

2.1 Thermostable trypsin derivates for enhanced in-gel digestion in high throughput proteomics

In collaboration with Prof. Dr. Marek Šebela from Department of Biochemistry, Palacky University, (Olomouc, Czech Republic) I tested trypsins conjugated with di-, tri-, tetrasaccharides and cyclodextrins in accelerated in-gel digestion of proteins as in protocol previously established in our laboratory [2]. These conjugates offer higher thermostability and autolysis-resistance compared to the commonly used in proteomics bovine trypsin. Their relatively small size represents a compromise between the stabilizing role of sugar moieties and molecular size of the enzyme.

2.1.1 Introduction in synthesis and bioanalytical characterization of bioconjugated enzymes

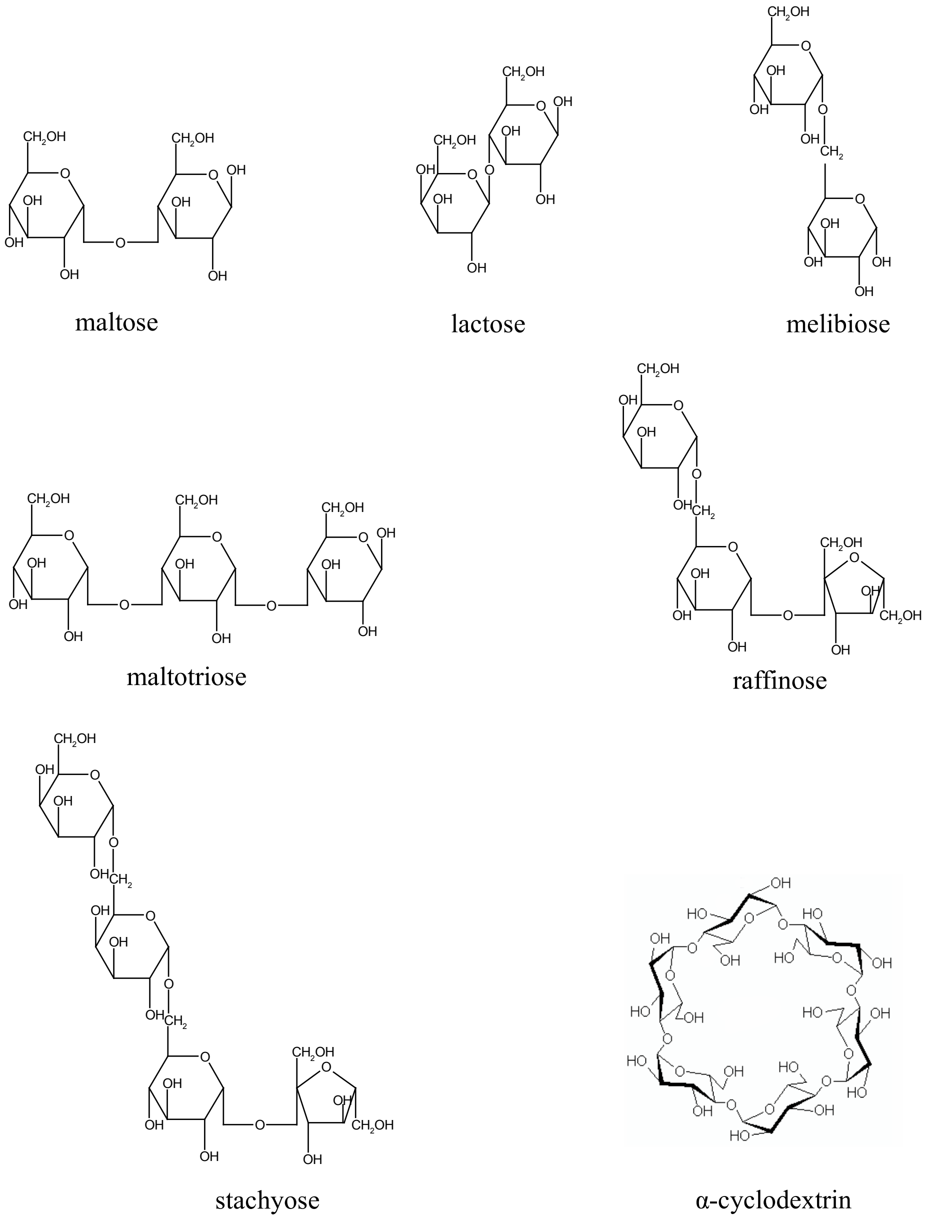
2.1.1.1 Chemical modification of bovine trypsin: glycosylation

Marek Šebela synthesized bovine trypsin conjugates by coupling oligosaccharides to its lysine residues (Figure 2.1).

Since the diffusion of enzymes into the gel pores during in-gel digestion is controlled by their size [2, 289], we selected oligosaccharides so, that if lysine residues were almost completely modified, the molecular mass of the conjugate should not exceed approximately 35 kDa [1].

To obtain lactose, maltose and melibiose trypsin conjugates (LAC-BT, MAL-BT and MEL-BT, respectively) trypsin was reacted directly by the aldehyde (acyclic) forms of the disaccharides in the presence of sodium cyanoborohydride, which reduced intermediate Schiff bases (Figure 2.2 A). Maltotriose, raffinose, stachyose, α- and β-cyclodextrin trypsin conjugates (MAT-BT, RAF-BT, STA-BT, ACD-BT and BCD-BT, respectively) were synthesized by coupling BT with oligosaccharides activated by potassium periodate oxidation (Figure 2.2 B). This was followed by the reduction with cyanoborohydride. Free arginyl residues in raffinose modified trypsin (RAF-BT) were optionally reacted with biacetyl, yielding an RAFR-BT with modified arginine residues.

In both glycosylation methods, BT was protected from autolysis during the reaction by its competitive inhibitor benzamidine.



maltose

lactose

melibiose

maltotriose

raffinose

stachyose

α-cyclodextrin

Figure 2.1. Oligosaccharides applied for chemical modification of bovine trypsin.

Disaccharides: maltose, lactose and melibiose; trisaccharides: maltotriose and raffinose; tetrasaccharide: stachyose and cyclodextrins (α-cyclodextrin).

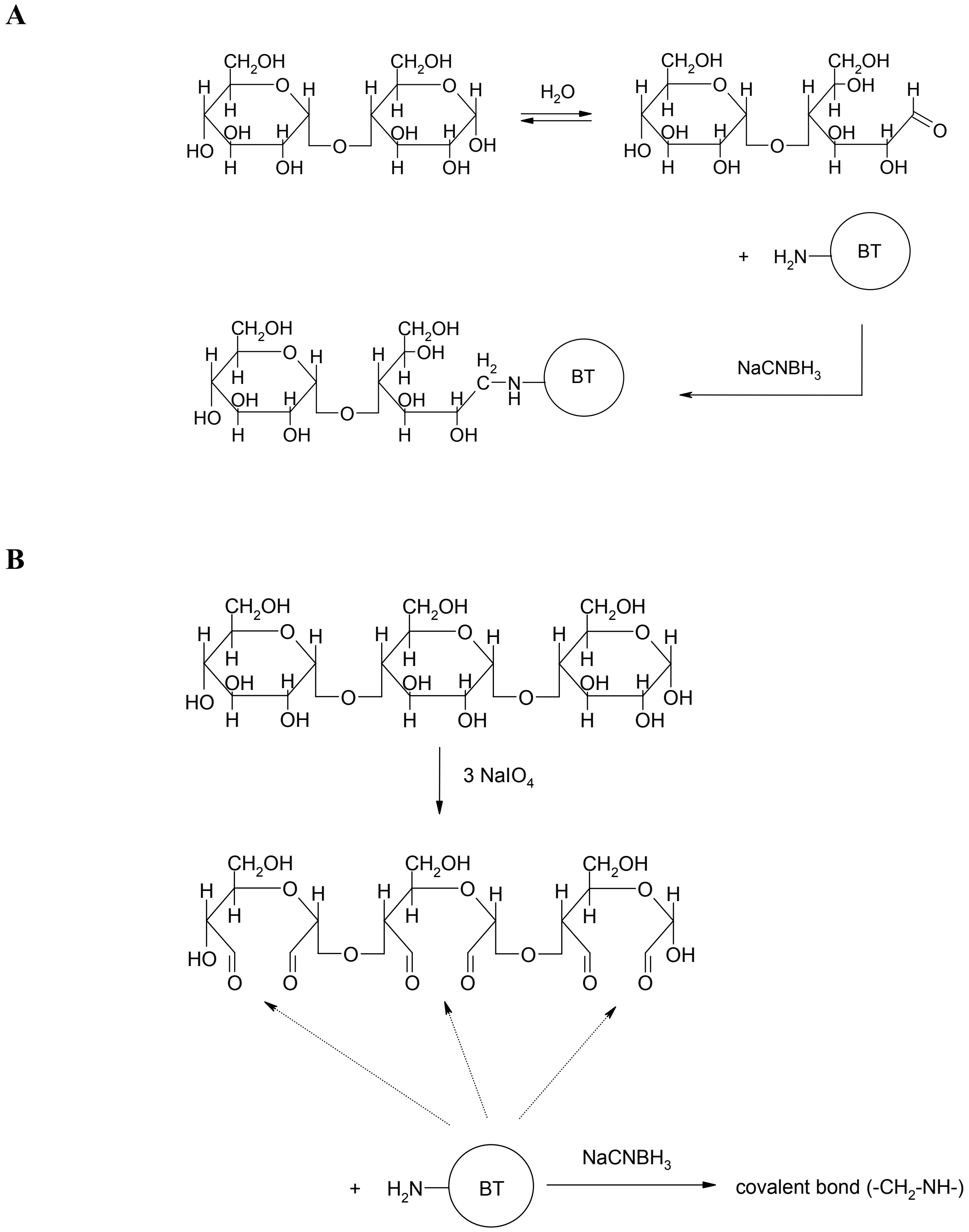


Figure 2.2. Preparation of saccharide modified trypsins.

A) Trypsin modification by disaccharides: a disaccharide (in this case, maltose) dissolved in water undergo mutarotation, resulting in partially acyclic molecules containing free aldehyde groups, which react with lysine residues of trypsin. The formed Schiff bases are subsequently reduced by cyanoborohydride.

B) Trypsin modification by higher oligosaccharides: an oligosaccharide (in this case, maltotriose) is first oxidized with sodium periodate, resulting in a polyaldehyde, which reacts with lysine residues of trypsin via formation of a Schiff base. The final stabilization is achieved by cyanoborohydride reduction.

Marek Šebela determined the extent of trypsin modification by oligosaccharides. The degree of saccharide modification was independently measured in three ways: by spectrophotometric quantification of free amino groups, by spectrophotometric quantification of neutral sugar content and, in some cases, amino acid analysis via the content of unmodified lysine residues determined by amino acid analysis.

Considering the total number of lysine residues in its sequence plus the N terminus, BT comprises 15 primary amino groups [141]. Despite large molar excess of modifying regents, a maximum of 9 modified lysine residues was achieved. The carbohydrate content of modified conjugates was in the range of 8 to 25%, which agreed well with the corresponding number of modified lysine residues. The amino acid analyses of RAF-BT, RAFR-BT and BCD-BT confirmed that the content of free lysine residues was substantially decreased (5, 4 and 5 residues per molecule, respectively) [1].

2.1.1.2 Glycosylated trypsins: molecular masses and pI values

To further characterize the obtained trypsin conjugates Marek Šebela determined their molecular masses and pI values.

The molecular masses of the conjugates were determined by discontinuous tricine-SDS-PAGE and by MALDI TOF MS. As anticipated, the molecular mass of the disaccharide conjugates of BT (~ 25 kDa) was only slightly higher than that of unmodified BT (~ 23 kDa), whereas masses of other conjugates were significantly higher (~ 27-33 kDa). The molecular mass of RAF-BT conjugate was directly determined by MALDI-TOF mass spectrometry as 26.29 kDa (Figure 2.3 A). Similarly, masses of MAL-BT and STA-BT were determined as 25.23 kDa and 28.52 kDa, respectively. Intact BT was detected as a narrow symmetric peak corresponding to a more accurate mass value of 23.29 kDa. After the coupling with relatively large molecules of the activated cyclodextrins, a strong mass heterogeneity of the produced BT conjugate was apparent. For example, MALDI-TOF spectrum of BCD-BT revealed a series of partially resolved peaks between m/z 29 032 and 33 260 with the most abundant component at m/z 31 126 (Figure 2.3 B). The mass differences between adjacent peaks in the series matched the mass of BCD [1].

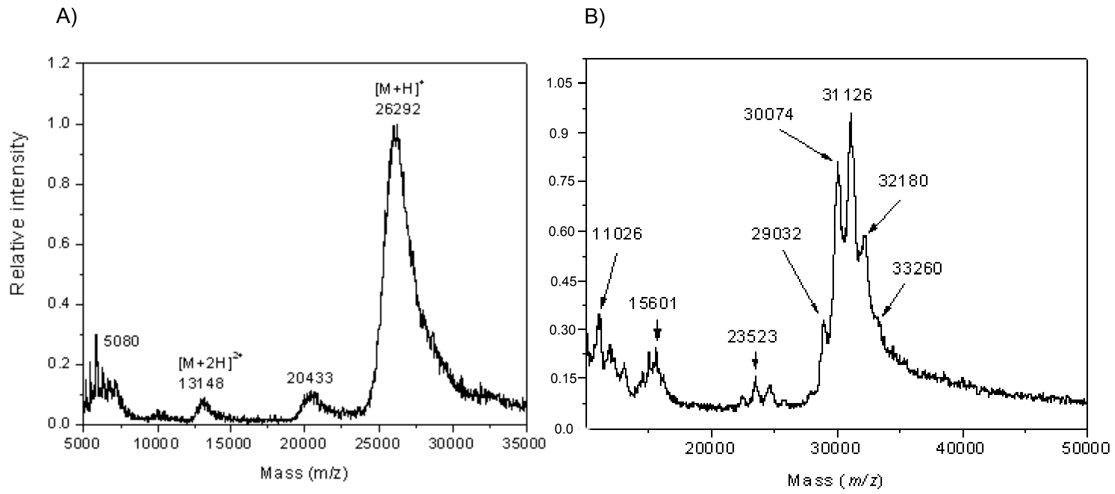


Figure 2.3. MALDI-TOF mass spectra of intact RAF-BT and BCD-BT.

(A) Spectrum of intact RAF-BT; (B) Spectrum of intact BCD-BT;
Spectra were acquired in the linear mode by Dr. Jan Havliš, from Laboratory of functional Genomics and proteomics, Masaryk University, Brno, Czech Republic.

The single-chain form of BT (β-trypsin) is a strongly basic protein with pI 10.5 [290]. Since lysine residues significantly contribute to the net charge, Šebela et al. performed IEF in order to estimate how their modification affected the pI of the enzyme. The RAF-BT band in IEF gel was more acidic (pI 6.7), compared to the native BT. Similar pI values were found for MAT-BT (6.6), RAFR-BT (5.9), STA-BT (6.3) and BCD-BT (6.1) [1].

2.1.1.3 *Activity and thermostability of glycosylated trypsins*

To characterize kinetic properties of the synthesized conjugates, Marek Šebela determined their specific activity and K_m values using a low molecular weight substrate BAPNA. The modification decreased the specific activity by 10-30 % compared to that of unmodified BT (28 nkat/mg), most substantially for the disaccharide conjugates (Table 2.1). Their K_m values were all in the millimolar range, with no considerable difference compared to unmodified BT (K_m = 2.8 mM). Thermostability of the conjugates was evaluated by their T_{50} constant, defined as a temperature at which 50% of the activity is retained upon 30 min incubation. T_{50} of unmodified BT was only 41°C. For LAC-BT, MAT-BT and MEL-BT, it was higher by ~ 10°C, and for MAT-BT, RAF-BT, RAFR-BT and STA-BT by ~ 20°C. Among all conjugates, ACD-BT and BCD-BT were the most stable with T_{50} close to 70°C [1] (Table 2.1).

Table 2.1: Catalytic activity and thermostability of saccharide modified trypsin conjugates determined by BAPNA substrate.

		native BT	LAC-BT	MAL-BT	MEL-BT	MAT-BT	RAF-BT	RAFR-BT	STA-BT	ACD-BT	BCD-BT
activity[a]	[nkat/mg]	28.4	10.7	11.4	11.6	25.6	21.4	18.1	20.6	22	20.5
K_m[b]	[mM]	2.8	3	2.8	2.8	3.6	3.1	3.1	5.1	3.3	4.2
T_{50}[c]	[°C]	41	50	48	49	60	57	68	58	67	68

[a] the calculated specific activity for BAPNA substrate;
[b] the calculated Michaelis constant for BAPNA substrate;
[c] the temperature at which 50% of enzyme activity is retained upon 30 min incubation;
The data were obtained by Prof. Dr. Marek Šebela, Department of Biochemistry, Palacky University, Olomouc, Czech Republic.

Figure 2.4 shows a plot of the residual enzyme activity versus the temperature of incubation for intact BT and for MAT-BT and BCD-BT conjugates.

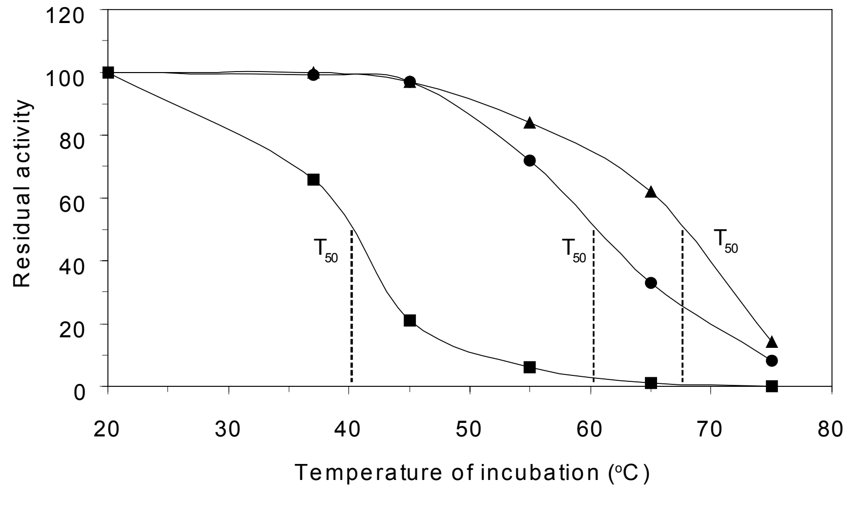


Figure 2.4. Thermostability of modified trypsin conjugates.

Enzyme aliquots were incubated at different temperatures ranging from 20 to 75°C for 30 min: BT (-■-), MAT-BT (-●-) and BCD-BT (-▲-). After rapid cooling, residual activity was determined by hydrolysis of BAPNA substrate at 30°C. The corresponding T_{50} values are indicated by vertical lines. The data were obtained by Prof. Dr. Marek Šebela, Department of Biochemistry, Palacky University, Olomouc, Czech Republic.

Importantly, the increased T_{50} resulted in more efficient cleavage of BAPNA at elevated temperatures. The rate of BAPNA cleavage by RAF-BT increased up to 55°C and then remained constant up to 70°C, whereas for BT it rapidly declined above 50°C [1].

2.1.1.4 Glycosylation of bovine trypsin: what was achieved?

Šebela et al. introduced oligosaccharide conjugating of bovine trypsin as facile and inexpensive method, which significantly increased its thermostability and suppressed autolysis. Since oligosaccharides moieties are relatively small the conjugates can be used for in-gel digestion of proteins. Better thermostability and autolysis resistance of glycosylated trypsins compared to its unmodified form enable their implementation in accelerated digestion protocol [2].

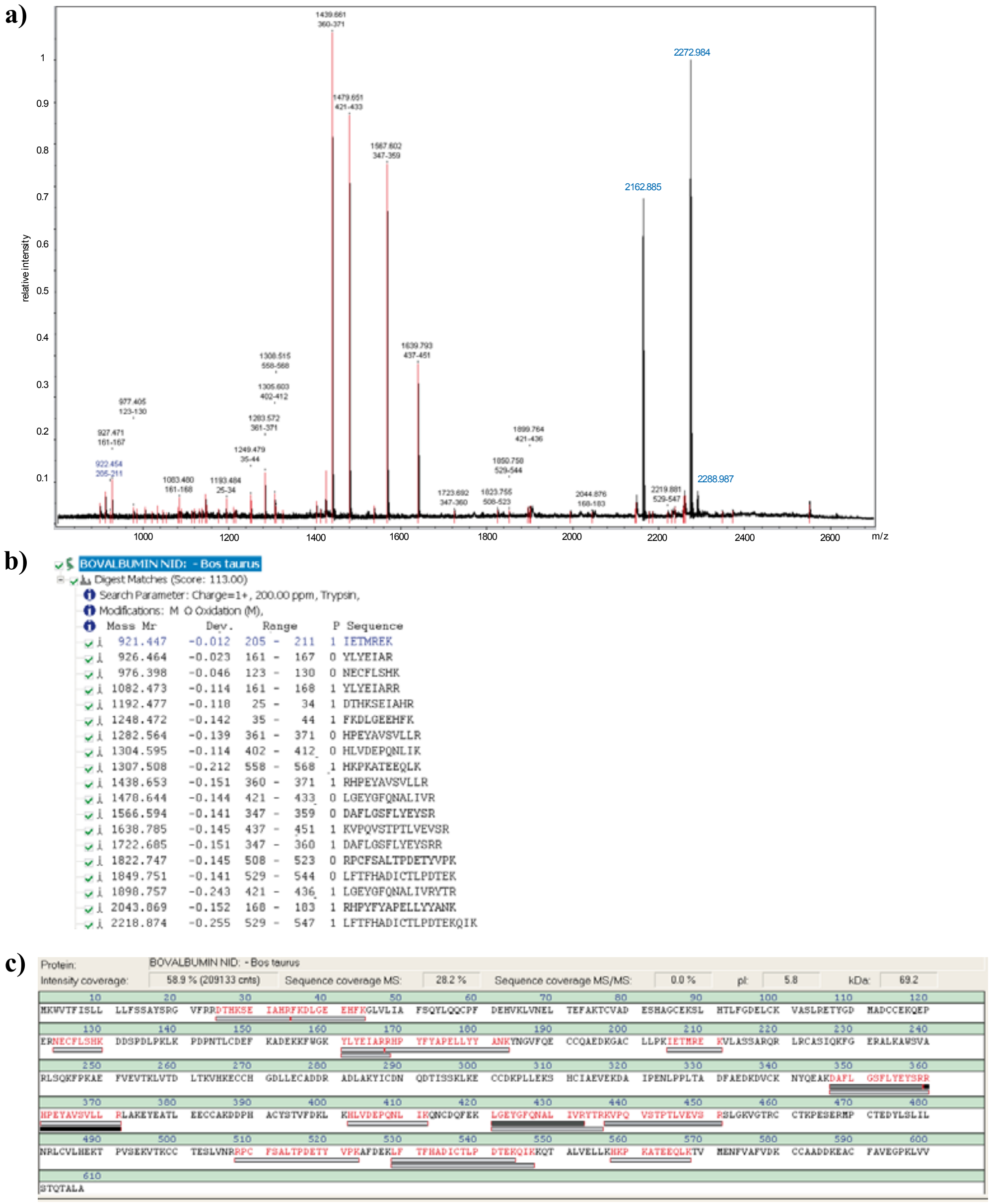
2.1.2 Performance of glycosylated trypsins in accelerated in-gel digestion of proteins

Conventional in gel-digestion by BT is performed at 37°C overnight. Accelerated in-gel digestion protocol (ADP) developed by Havliš et al. applies thermostable methylated porcine trypsin at higher enzyme concentrations (compared to conventional digestion) and at higher temperature (55°C), enabling to reduce digestion time to 0.5-1 h [2]. Based on the kinetic study Havliš demonstrated that ADP dramatically simplifies and accelerates the sample preparation routine without compromising the yield of digestion products, sensitivity of peptide detection and confidence of protein identification. In line with protocol established by Havliš et al. [2] I set out to evaluate the performance of trypsin conjugates in in-gel digestion at accelerated temperature. To this end I digested in-gel several standard proteins using BT (under conventional conditions) and its glycosylated conjugates (under accelerated conditions). The obtained digests were analyzed by MALDI TOF MS. Here I aimed to compare the quality of their peptide mass fingerprints and confidence of protein identification. Further I aimed to analyze whether the cleavage specifity of the modified enzymes was altered compared to their unmodified form. And finally I compared the number of autolytic products and their abundance of trypsin conjugates with those of unmodified bovine trypsin.

2.1.2.1 In-gel digestion of protein standards by glycosylated trypsins

5 pmol of standard proteins (Cytochrom C, Myoglobin, Aldolase, and BSA) were separated by gel electrophoresis, stained with Coomasie and digested by BT (under conventional conditions: overnight digestion at 37°C and enzyme concentration about 0.5 µM) and glycosylated trypsins (under accelerated conditions: 1-3 h digestion at 55°C and higher enzyme concentration ranging from 0.9 to 3 µM). The obtained peptides were subsequently analyzed by MALDI-TOF MS.

The reaction temperature of 55°C was selected to balance the reaction rate against the rate of thermal inactivation, both of which accelerate along with the temperature increase. The relatively high load of protein standards allowed me to acquire spectra that were rich in tryptic peptides, and hence I could better evaluate and find possible changes in the cleavage specificity of the trypsin conjugates. In acquired MALDI TOF spectra, *m/z* of all peaks with S/N ratio > 2 were fetched and used for searches against MSDB protein sequence database with mass tolerance 150 - 200 ppm. Figure 2.5 a represents peptide mass fingerprint of a BSA in-gel digest (5 pmol) obtained by accelerated digestion using RAF-BT at 55°C and at the enzyme concentration of 0.86 µM; the digestion time was 1.5h. All abundant peaks matched m/z of BSA tryptic peptides when searched against protein sequence database, whereas autolytic background (see chapter 2.1.2.2) of the applied enzyme (peaks m/z 2162.885, 2272.984 and 2288.987, corresponding to the autolysis products LGEDNINVVEGNEQFISASK, SIVHPSYNSNTLNNDIMLIK and SIVHPSYNSNTLNNDIM(ox)LIK) didn’t overload the spectrum and didn’t hamper the peak picking. Altogether, nineteen BSA peptides could be identified upon database search (Figure 2.5 b), covering 28% of protein sequence (Figure 2.5 c).



```
BOVALBUMIN NID: - Bos taurus
Digest Matches (Score: 113.00)
Search Parameter: Charge=1+, 200.00 ppm, Trypsin,
Modifications: M O Oxidation (M),
  Mass Mr       Dev.    Range      P Sequence
   921.447    -0.012   205 -  211  1 IETMREK
   926.464    -0.023   161 -  167  0 YLYEIAR
   976.398    -0.046   123 -  130  0 NECFLSHK
  1082.473    -0.114   161 -  168  1 YLYEIARR
  1192.477    -0.118    25 -   34  1 DTHKSEIAHR
  1248.472    -0.142    35 -   44  1 FKDLGEEHFK
  1282.564    -0.139   361 -  371  0 HPEYAVSVLLR
  1304.595    -0.114   402 -  412  0 HLVDEPQNLIK
  1307.508    -0.212   558 -  568  1 HKPKATEEQLK
  1438.653    -0.151   360 -  371  1 RHPEYAVSVLLR
  1478.644    -0.144   421 -  433  0 LGEYGFQNALIVR
  1566.594    -0.141   347 -  359  0 DAFLGSFLYEYSR
  1638.785    -0.145   437 -  451  1 KVPQVSTPTLVEVSR
  1722.685    -0.151   347 -  360  1 DAFLGSFLYEYSRR
  1822.747    -0.145   508 -  523  0 RPCFSALTPDETYVPK
  1849.751    -0.141   529 -  544  0 LFTFHADICTLPDTEK
  1898.757    -0.243   421 -  436  1 LGEYGFQNALIVRYTR
  2043.869    -0.152   168 -  183  1 RHPYFYAPELLYYANK
  2218.874    -0.255   529 -  547  1 LFTFHADICTLPDTEKQIK
```

```
Protein:             BOVALBUMIN NID: - Bos taurus
Intensity coverage:  58.9 % (209133 cnts)   Sequence coverage MS: 28.2 %   Sequence coverage MS/MS: 0.0 %   pI: 5.8   kDa: 69.2

        10         20         30         40         50         60         70         80         90        100        110        120
MKWVTFISLL LLFSSAYSRG VFRRDTHKSE IAHRFKDLGE EHFKGLVLIA FSQYLQQCPF DEHVKLVNEL TEFAKTCVAD ESHAGCEKSL HTLFGDELCK VASLRETYGD MADCCEKQEP
       130        140        150        160        170        180        190        200        210        220        230        240
ERNECFLSHK DDSPDLPKLK PDPNTLCDEF KADEKKFWGK YLYEIARRHP YFYAPELLYY ANKYNGVFQE CCQAEDKGAC LLPKIETMRE KVLASSARQR LRCASIQKFG ERALKAWSVA
       250        260        270        280        290        300        310        320        330        340        350        360
RLSQKFPKAE FVEVTKLVTD LTKVHKECCH GDLLECADDR ADLAKYICDN QDTISSKLKE CCDKPLLEKS HCIAEVEKDA IPENLPPLTA DFAEDKDVCK NYQEAKDAFL GSFLYEYSRR
       370        380        390        400        410        420        430        440        450        460        470        480
HPEYAVSVLL RLAKEYEATL EECCAKDDPH ACYSTVFDKL KHLVDEPQNL IKQNCDQFEK LGEYGFQNAL IVRYTRKVPQ VSTPTLVEVS RSLGKVGTRC CTKPESERMP CTEDYLSLIL
       490        500        510        520        530        540        550        560        570        580        590        600
NRLCVLHEKT PVSEKVTKCC TESLVNRRPC FSALTPDETY VPKAFDEKLF TFHADICTLP DTEKQIKKQT ALVELLKHKP KATEEQLKTV MENFVAFVDK CCAADDKEAC FAVEGPKLVV
       610
STQTALA
```

Figure 2.5: MALDI TOF MS analysis and protein identification upon database searching of BSA in-gel digest obtained by accelerated digestion using RAF-BT.

(a) Peptide mass map of a BSA in-gel digest. A gel band containing 5 pmol BSA was digested by RAF-BT at 55°C; the digestion time was set at 1.5h and the applied enzyme concentration was 0.86 µM (enzyme stock solution was subjected to amino acid analysis in the laboratory of Dr. Hunziker (University of Zürich, Switzerland)); (b) peptides identified upon database searching. Peaks with S/N ratio > 2 were fetched and submitted for searches against MSDB protein sequence database with mass tolerance 150 - 200 ppm, up to one miss cleavage site was allowed and oxidation of methionine was considered as possible modification. BSA tryptic peptides are highlighted by red lines in the spectrum; peaks 2162.885, 2272.984 and 2288.987 correspond to the tryptic peptides of RAF-BT: LGEDNINVVEGNEQFISASK, SIVHPSYNSNTLNNDIMLIK and SIVHPSYNSNTLNNDIM(ox)LIK; (c) sequence coverage of identified peptides.

Further I compared the sequence coverage (determined as % of the full-length protein sequence covered with the matched peptides) of peptide mass maps of the digests produced by the BT conjugates with the maps obtained by conventional digestion using BT (37°C, overnight) or accelerated digestion using MET-PT (1 h at 55°C) [2] (Table 2.2).

MAT-BT, RAF-BT, RAFR-BT and STA-BT performed well under conditions of the accelerated in-gel digestion protocol [2]. The sequence coverage and MOWSE scores [170] (a merit of statistical significance provided by MASCOT database searching software) of the peptide mass maps acquired from the digests with glycosylated trypsins or MET-PT (1h at 55°C) and with BT (overnight, 37°C) were similar [1] (Table 2.2). Moreover, MALDI TOF peptide mass fingerprints obtained from digests by trypsin conjugates were comparable to those acquired from digests by unmodified BT or by methylated porcine trypsin, suggesting that their cleavage specifity remained unchanged.

Table 2.2: MALDI TOF peptide mass fingerprints of protein standards in-gel digested by BT and its conjugates.

Protein standard	BT (Roche)		MET-PT (Promega)		MAT-BT		RAF-BT		RAFR-BT		STA-BT	
	Peptides	Coverage (%)	Peptides	Coverage (%)	Peptides	Coverage (%)	Peptides	Coverage (%)	Peptides	Coverage (%)	Peptides	Coverage (%)
Cytochrome C	7	53	10	63	7	30	5	45	6	45	7	43
Myoglobin	12	76	10	71	11	74	11	74	11	74	12	80
Aldolase	15	34	19	49	15	46	15	39	17	45	17	47
BSA	20	32	21	35	16	27	19	28	14	20	10	17

Applied digestion conditions: conventional digestion was performed overnight at 37°C by commercially available BT (Roche) at an enzyme concentration of 0.5 μM; accelerated digestion by MET-PT (Promega) was performed for 1.5 h at 55°C and an enzyme concentration of 0.75 μM; accelerated digestion by glycosylated trypsins was performed for 1.5 h at 55°C; the applied enzyme concentrations were for MAT-BT 0.98 μM, for RAF-BT 0.86 μM, for RAFR-BT 1.2 μM and for STA-BT 1.04 μM.

In MALDI-TOF spectra of in-gel digest obtained by LAC-BT, MAL-BT and MEL-BT conjugates none detectable peptides of the analyzed proteins were found. ACD-BT and BCD-BT, which efficiently digested proteins in solution, demonstrated only marginal activity in in-gel digestion. Figure 2.6 demonstrates the peptide yield observed by in-gel digestion of BSA using methylated porcine trypsin and glycosylated trypsins under accelerated conditions as described in experiment before. Digestion yields were obtained in kinetic study (see chapter 2.1.3) using ^{18}O-labeled peptides as internal standards. The determined yield of conventional digestion was 2.8 pmol, accelerated digestion by MET-PT achieved 76 % of CDP yield, whereas MAT-BT, RAF-BT, RAFR-BT, STA-BT and BCD-BT reached 33, 39, 63, 20 and 5 % of CDP yield, respectively. Although the efficiency of glycosylated trypsins is lower than those of methylated porcine trypsin, their higher thermostability and autolysis resistance enable adjustment of optimal digestion conditions at higher temperature and enzyme concentration.

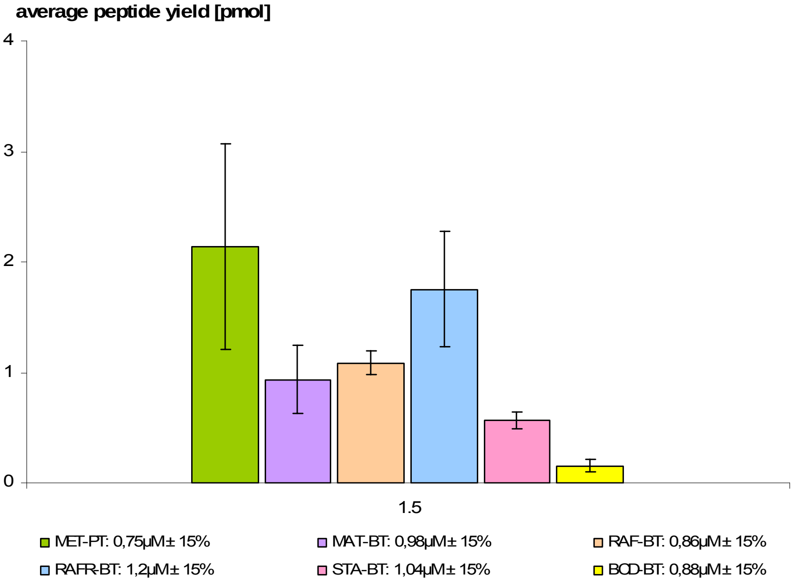


Figure 2.6: Bar diagram representing averaged peptide yield observed by in-gel digestion of BSA using methylated porcine trypsin and glycosylated trypsin derivates under accelerated conditions.

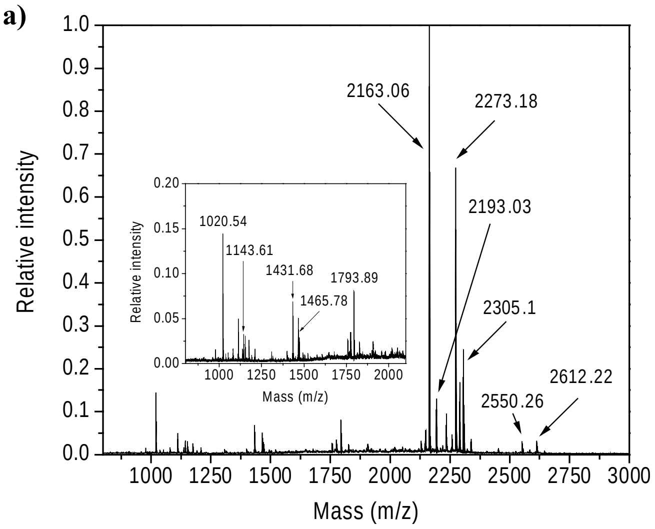
Gel bands containing 5 pmol BSA were digested by MET-PT, MAT-BT, RAF-BT, RAFR-BT, STA-BT and BCD-BT at 55°C; the digestion time was set at 1.5h and the applied enzyme concentrations were 0.75; 0.98; 0.86; 1.2; 1.04 and 0.88 µM, respectively. The digestion yields were obtained by kinetic study using ^{18}O-labeled peptides as internal standards (as described in chapter 2.1.3). The coloured bars represent the averaged digestion yields generated by MET-PT, MAT-BT, RAF-BT, RAFR-BT, STA-BT and BCD-BT (green, purple, pale pink, blue, pink and yellow, respectively). The digestion of CDP was 2.8 pmol.

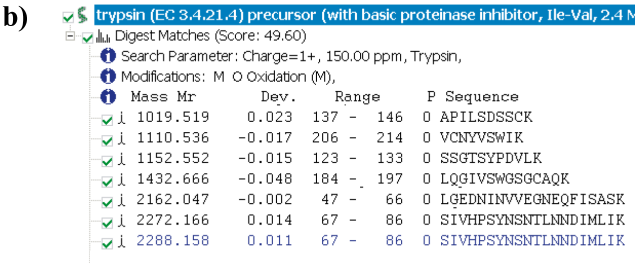
Since the digestion efficiency of BCD and ACD-BT conjugates was very low only MAT-BT, RAF-BT, RAFR-BT and STA-BT were subjected to further kinetic study.

2.1.2.2 Autolytic background of glycosylated trypsins

To determine the number and relative abundance of autolysis products, I performed control digests of blank gel slabs. The digestion was performed overnight at 37°C and at an enzyme concentration of ~ 1.0 μM. MALDI-TOF mass spectra of autodigests of MAT-BT, RAF-BT, RAFR-BT and STA-BT were acquired and compared with the autolytic peptide pattern of BT. Among detected peaks 1020.54 (peptide APILSDSSCK), 1111.49 (peptide VCNYVSWIK), 2163.06 (peptide LGEDNINVVEGNEQFISASK), 2273.18 (peptide SIVHPSYNSNTLNNDIMLIK) and 2289.18 (peptide SIVHPSYNSNTLNNDIM(ox)LIK) were major autolytic peaks of unmodified bovine trypsin (Fig. 2.7 a). But also peptides originating from human keratins and minor autolysis products of trypsin were detected. Altogether, six tryptic peptides of BT were identified by search in the MSDB protein sequence database (Fig. 2.7 b). On the contrary, only three tryptic peptides from BT (Figure 2.8), corresponding to the peptides LGEDNINVVEGNEQFISASK with m/z 2163.1, SIVHPSYNSN-TLNNDIMLIK with m/z 2273.2 and SIVHPSYNSNTLNNDIM(ox)LIK with m/z 2289.2 were found in the autodigests of MAT-BT, RAF-BT, RAFR-BT and STA-BT (Fig. 2.7 b). The intensity ratio of the major peaks (m/z 2163.1 and 2273.2) strongly varied among the spectra. Importantly, the autodigests of the conjugates contained fewer minor autolytic peptides, which complicate database searching and might lead to incorrect interpretations. Thus, lower number of autolytic and background peptides of trypsin conjugates enhanced the specifity of protein identification (Fig. 2.7 b).

Interestingly, that the peaks at m/z 2163 and 2273 detected in autodigests of BT (as the most abundant peaks) and its conjugates contain C-terminal Lys but not Arg residues. The sequence of bovine trypsinogen (Swiss-Prot access. code P00760) comprises 243 amino acids, from which the region 21-243 represents the mature chain of β-trypsin. The above peptides are located in successive positions 70-89 and 90-109. The crystal structure of BT complex with 2-aminobenzimidazole) was downloaded from the RCSB Protein Data Bank (www.rcsb.org/pdb). Using the program DeepView/Swiss-PdbViewer v3.7 (www.expasy.org/spdbv), we observed that Lys89 and Lys109 are located close to each other at the molecule surface being in a distance of 7 Å (Figure 2.9). Because of sterical reasons, they probably cannot be both reacted by bulky substituents (for example α-cyclodextrin molecule is about 20 Å long [291], but there is only one of them modified randomly [1].





```
trypsin (EC 3.4.21.4) precursor (with basic proteinase inhibitor, Ile-Val, 2.4 M
 Digest Matches (Score: 49.60)
   Search Parameter: Charge=1+, 150.00 ppm, Trypsin,
   Modifications:  M  O Oxidation (M),
    Mass Mr        Dev.     Range      P Sequence
   1019.519     0.023   137 -   146   0 APILSDSSCK
   1110.536    -0.017   206 -   214   0 VCNYVSWIK
   1152.552    -0.015   123 -   133   0 SSGTSYPDVLK
   1432.666    -0.048   184 -   197   0 LQGIVSWGSGCAQK
   2162.047    -0.002    47 -    66   0 LGEDNINVVEGNEQFISASK
   2272.166     0.014    67 -    86   0 SIVHPSYNSNTLNNDIMLIK
   2288.158     0.011    67 -    86   0 SIVHPSYNSNTLNNDIMLIK
```

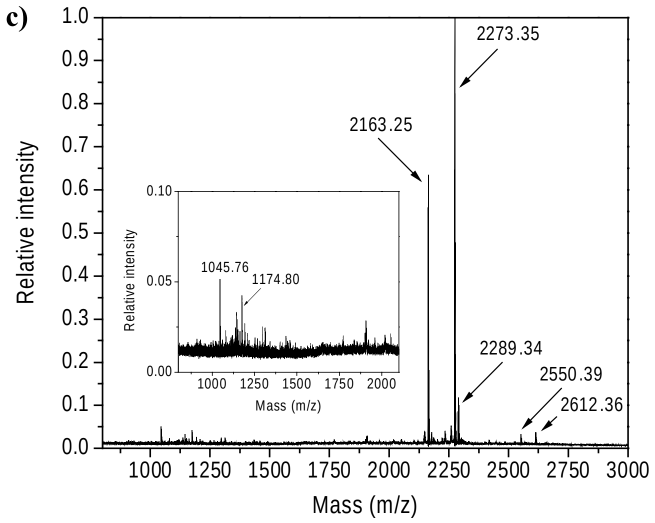


Figure 2.7: Peptide mass fingerprints of autolysis products of BT and MAT-BT.

(a) Peptide mass fingerprint of BT autolyzate; (b) trypsin peptides identified from the spectrum (a) by MASCOT search against MSDB protein sequence database with mass tolerance of 150 ppm (c) peptide mass fingerprint of MAT-BT autolyzate. Blank gel slabs were incubated in BT and MAT-BT (both 1.0 µM) in 50 mM ammonium bicarbonate at 37°C for 12 h. Then aliquots (1 µL) were withdrawn and analyzed by MALDI-TOF MS using a CHCA matrix.

```
GYTCGANTVPYQVSLNSGYHFCGGSLINSQWVVSAAHCYKSGIQVR  LGEDNINVVEGNEQFISASK  SIVHPSYNSNTLNNDIMLIK  LKSAASLNSRVASISLPTSCASAGTQCLISGWGN
                                               ----------BT---------  --------BT----------
                                               --------RAF-BT-------  -------RAF-BT-------
                                               --------RAFR-BT------  -------RAFR-BT------
                                               --------STA-BT-------  -------STA-BT-------
                                               --------BCD-BT-------  -------BCD-BT-------

TK  SSGTSYPDVLK  CLK  APILSDSSCK  SAYPGQITSNMFCAGYLEGGKDSCQGDSGGPVVCSGK  LQGIVSWGSGCAQK  NKPGVYTK  VCNYVSWIK  QTIASN
    ----BT-----       ----BT----                                         -----BT-------            ---BT----
```

Figure 2.8: Autolytic peptides of BT and its conjugates within sequence of BT.

Tryptic peptides of BT and its conjugates detected by MALDI TOF MS are highlighted in red colour. Dotted lines underline peptides obtained by autolysis of BT or of its glycosylated derivates.

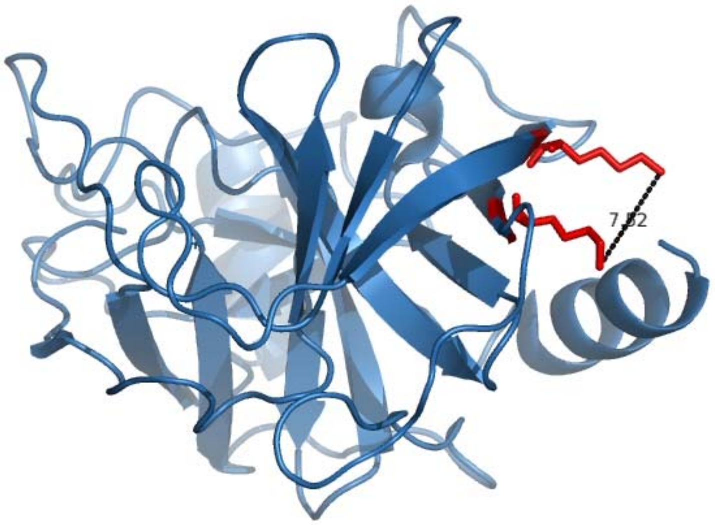


Figure 2.9: Crystal structure of BT complex with 2-aminobenzimidazole.

The picture was downloaded from the RCSB Protein Data Bank (www.rcsb.org/pdb). Program DeepView/Swiss-PdbViewer v3.7 (www.expasy.org/spdbv) was used to observe location of trypsin cleavage sites. Red lines represent peptides LGEDNINVVEGNEQFISASK (m/z 2163.1) and SIVHPSYNSNTLNNDIMLIK (m/z 2273.2); dotted line shows difference of 7 Å between Lys89 and Lys109.

2.1.2.3 Dried-droplet probe preparation method for MALDI analysis

For preparation of MALDI probes I applied dried-droplet method developed by Thomas et al. [292]. In this method peptides retain and co-crystallize with the CHCA matrix at the hydrophobic polymer surface of the target, while salts and the hydrophilic impurities are pooled at the hydrophilic metal anchor. Concentration of the matrix, as well as of water and organic solvent are perfectly adjusted in the dried-droplet method for MALDI TOF MS analysis of the tryptic digests produced by CDP, resulting in high sensitivity, low matrix-related background and high quality of the acquired spectra. However, this method has been observed to be less compatible with saccharide modified trypsins. Matrix crystals obtained from in-gel digests of proteins by glycosylated trypsins were visually normal (comparable with these derived from conventional digestion), but they were relatively rapidly depleted by laser pulses and often produced low quality noisy spectra. From several crystals no peptide signals were detected in the acquired TOF mass spectra. These effects were stronger pronounced at higher enzyme concentration (above 1.5 μM), suggesting that sugar oligomers might inhibit desorption of peptides from the matrix crystals (since no trypsin autolysis products were also observed). In addition, presence of hydrophilic sugar residues might change crystallization efficiency of the tryptic peptides on the target.

In order to achieve acquisition of good quality spectra accumulation of high number of shots was in all experiment required.

2.1.2.4 Performance of glycosylated trypsins in accelerated in-gel digestion of proteins: what did we learn?

In this part of my work I demonstrated that glycosylated trypsins efficiently digest proteins under accelerated conditions. They have the same cleavage specificity as BT and produce less autolytic background, hence increasing the specifity of protein identification. Their better thermostability and autolysis resistance compared to MET-PT make them promising candidates to further improve protocol of accelerated in-gel digestion of proteins developed by Havliš et al. [2].

Dried-droplet probe preparation routinely applied for MALDI analysis of tryptic digests obtained from CDP was less compatible with digests produced by saccharide modified trypsins. Generally, increased acquisition time should be taken in account in order to acquire spectra of sufficient quality. The above described difficulties in acquisition of spectra at increased concentration of glycosylated trypsins set a limit for the applied enzyme concentration in further experiments.

2.1.3 Kinetic study of accelerated in-gel digestion of proteins by glycosylated trypsins

Next I set out to evaluate the catalytic efficiency of trypsin conjugates in accelerated in-gel digestion of proteins. To this end I studied kinetics of in-gel digestion by MALDI TOF MS and applied ^{18}O-labeled peptides as internal standards for quantifying the yield of digestion products. In order to optimize the digestion conditions, I aimed to study the effect of the temperature, enzyme concentration and digestion time on the digestion yield. Optimized accelerated digestion protocol was subsequently applied by the identification of members of a protein complex isolated from the budding yeast.

2.1.3.1 Quantification method: ^{18}O labeling and deconvolution

The study of digestion kinetics relies on quantifying the yield of in-gel digestion products for optimization of digestion efficiency. A relatively simple and convenient isotope labeling approach is based on endoprotease-catalyzed incorporation of ^{18}O atoms in the C-terminal carboxylic acids during digestion of proteins [259, 293, 294].

^{18}O-labeled internal standards applied in my quantification experiments were generated by in-solution digestion of the model protein BSA in isotopically enriched water containing 95% ^{18}O. Previously Havliš et al. evaluated the yield of in-solution digestion using synthetic peptides and showed that it is close to 100% [2]. Therefore, the protein concentration should be directly proportional to the average concentration of individual tryptic peptides in the digest. Thus ^{18}O-labeled peptides can be used for absolute quantification of digestion products.

The general scheme of quantification experiment by ^{18}O-labeled internal standards is depicted in Figure 2.10. BSA was used as a model protein for the kinetic study. The amount of protein contained in a gel band was relatively high (5 pmol), enabling better signal-to-noise ratio, which improves the accuracy of quantification. In addition, it allowed compensating various yields of digestion products generated by differently modified conjugates.

In-gel digestion was performed as described in chapter 4.1.2.4 according to the tested digestion conditions. Obtained peptides were extracted from the gel matrix and extracts were subsequently dried down in a vacuum centrifuge. To prepare a mixture of

^{18}O-labeled peptides for the quantification experiment, a solution of 0.3 pmol/µL BSA in 25 mM ammonium bicarbonate buffer in $H_2{}^{18}O$ was digested overnight at 37°C and an enzyme:substrate ratio 1:50 (w/w). Tryptic peptides from in-gel digests were redissolved in 10 µL of internal standard and analyzed by MALDI TOF MS.

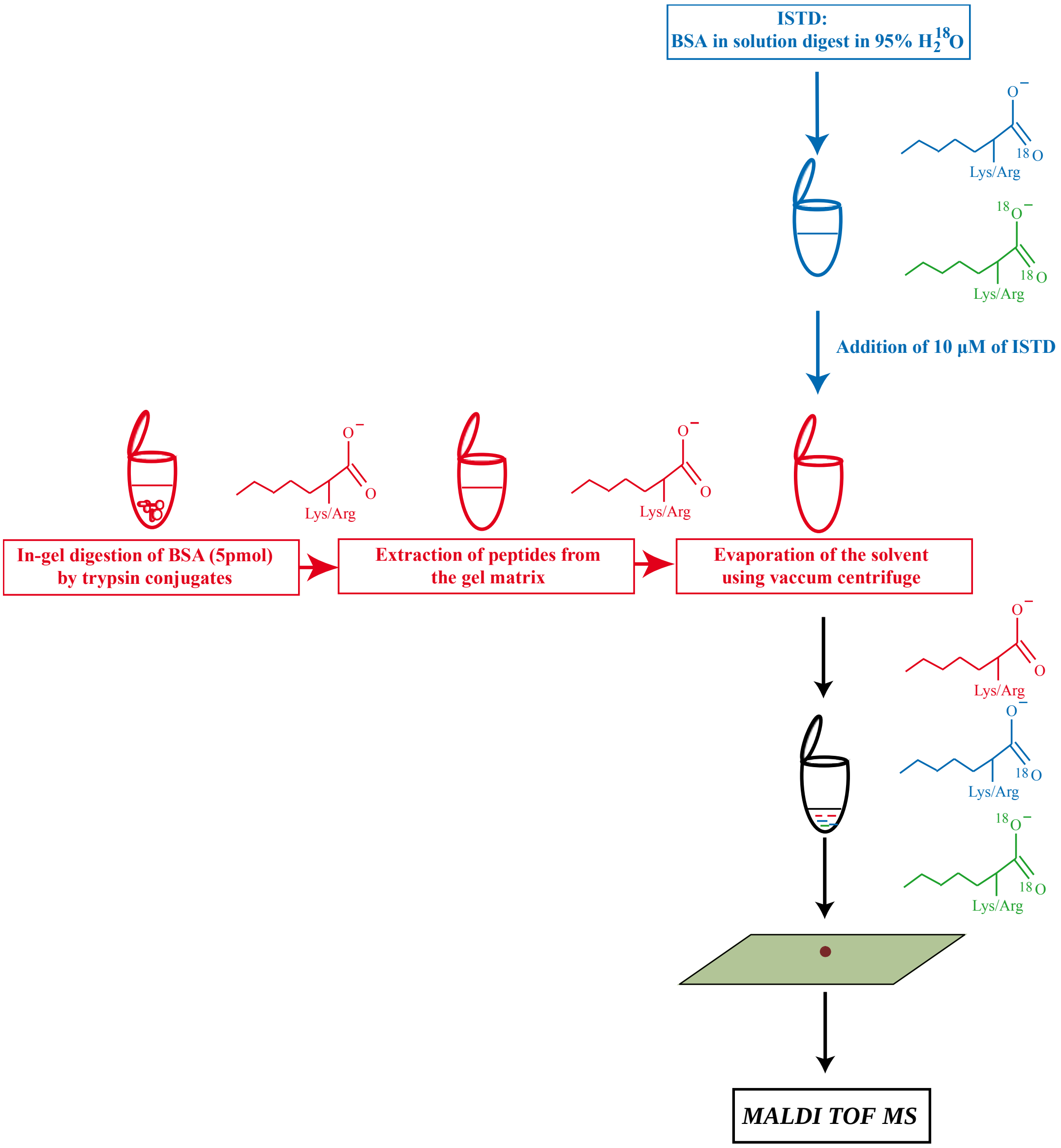
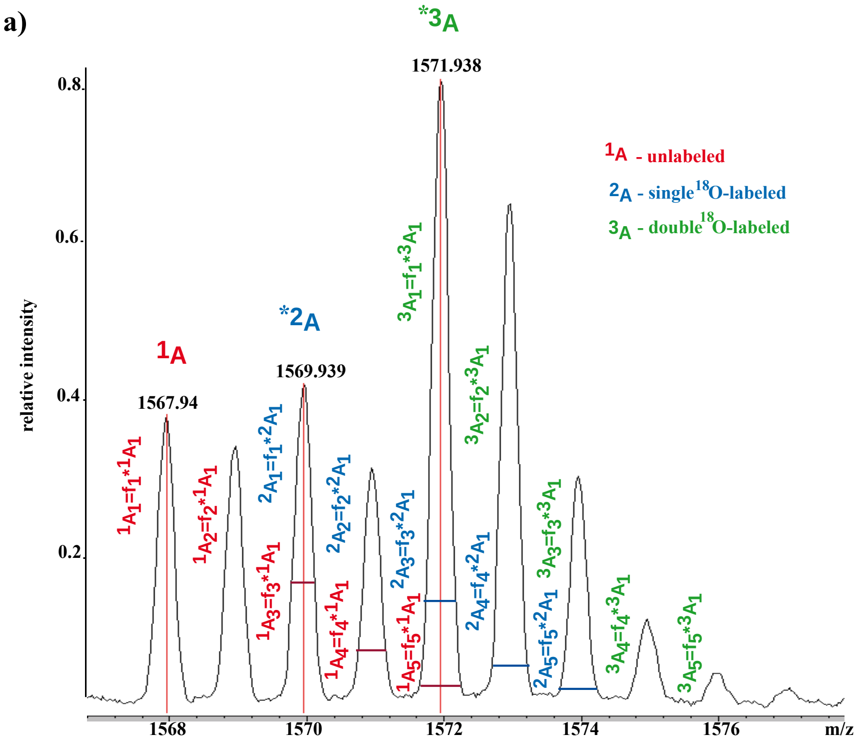
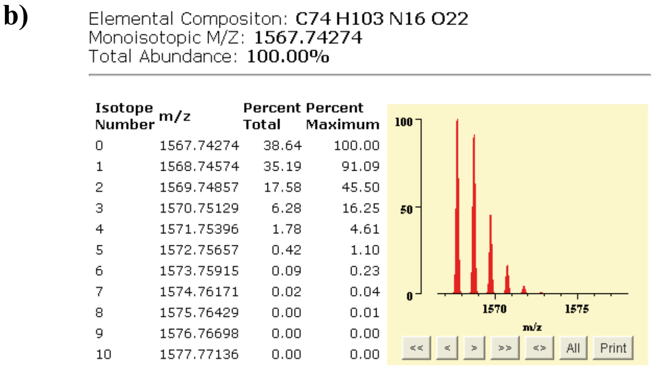

ISTD:
BSA in solution digest in 95% $H_2^{18}O$
O^-
^{18}O
Lys/Arg
$^{18}O^-$
^{18}O
Lys/Arg
Addition of 10 µM of ISTD
O^-
O
Lys/Arg
O^-
O
Lys/Arg
In-gel digestion of BSA (5pmol) by trypsin conjugates
Extraction of peptides from the gel matrix
Evaporation of the solvent using vaccum centrifuge
O^-
O
Lys/Arg
O^-
^{18}O
Lys/Arg
$^{18}O^-$
^{18}O
Lys/Arg
MALDI TOF MS

Figure 2.10: A workflow for absolute quantification of in-gel digestion products using ^{18}O-labeled peptides as internal standards.

Part highlighted in red colour represents the workflow of in-gel digestion; BSA (5pmol) was in-gel digested according to the tested conditions; the tryptic peptides were extracted and the extract was dried down. Subsequently the peptides were redissolved in 10 µL of 0.3 µM ISTD obtained by tryptic digestion of BSA in the buffer containing 95% $H_2{}^{18}O$. An aliquot from the obtained mixture was withdrawn, cocrystallized on a sample plate with a matrix solution and analyzed by MALDI TOF MS.

The incorporation of second ^{18}O into the carboxyl termini of peptides is usually incomplete in enzyme-catalyzed reaction. Figure 2.11 a shows merged isotopic clusters containing unlabeled BSA peptide DAFLGSFLYEYSR (m/z 1567.74) and its mono-^{18}O-and double-^{18}O-labeled internal standard.





Elemental Compositon: C74 H103 N16 O22
Monoisotopic M/Z: 1567.74274
Total Abundance: 100.00%

Isotope Number	m/z	Percent Total	Percent Maximum
0	1567.74274	38.64	100.00
1	1568.74574	35.19	91.09
2	1569.74857	17.58	45.50
3	1570.75129	6.28	16.25
4	1571.75396	1.78	4.61
5	1572.75657	0.42	1.10
6	1573.75915	0.09	0.23
7	1574.76171	0.02	0.04
8	1575.76429	0.00	0.01
9	1576.76698	0.00	0.00
10	1577.77136	0.00	0.00

Figure 2.11: Spectral pattern of merged isotopic clusters of a BSA peptide DAFLGSFLYEYSR (m/z 1567.74) and its ^{18}O-labeled standard.

a) Merged isotopic clusters containing unlabeled peptide and its mono-^{18}O-and double-^{18}O-labeled internal standard. Symbol A represents peak areas. ^{1}A (highlighted in red) labeled peaks refer to underivatized peptide isotopic peaks ($^{1}A_1$ – $^{1}A_5$), ^{2}A (highlighted in blue) refer to single-^{18}O-labeled peptide ($^{2}A_1$ – $^{2}A_5$) and ^{3}A (highlighted in green) refer to double-^{18}O-labeled peptide (peaks ($^{3}A_1$ – $^{3}A_5$). The star character refers to the convoluted peak area value (e.g. $^{*3}A = {}^{3}A_1 + {}^{1}A_5 + {}^{2}A_3$). (b) Isotopic distribution for peptide DAFLGSFLYEYSR (m/z 1567.74) computed by program PeptideProspector 4.0.4 (University of California, http://prospector.ucsf.edu); coefficients f_1-f_5 were: 1; 0.91; 0.46; 0.16; 0.05, respectively.

Monoisotopic peaks from single-^{18}O-labeled and double-^{18}O-labeled internal standard differ from the monoisotopic peak of unlabeled peptide by 2 and 4 Da, respectively. In Figure 2.11 a the monoisotopic peak at m/z 1569.94 of the single-^{18}O-labeled internal standard overlaps with third isotopic peak of the unlabeled peptide (45.5 % of the intensity of the monoisotopic peak). The intensity of the monoisotopic peak m/z 1571.94 of the double-^{18}O-labeled internal standard is affected by third isotopic peak of the single-^{18}O-labeled peptide and by fifth isotopic peak of the unlabeled peptide.

To calculate the peptide amount from MALDI TOF spectra a signal deconvolution method was employed.

The relation between the amount of non-labeled peptide of a sample, n_{16}, and both forms of ^{18}O-labeled peptide amount of the internal standard, n_{18} is defined in equation 1. The A_{16}/A_{18} represents the ratio of peak areas of sample and internal standard.

$$n_{16} = \frac{A_{16}}{A_{18}} \cdot n_{18}$$

(Equation 1)

The equation 2 defines the peak areas for both, sample and internal standard, presuming that maximum of 5 isotopic peaks per compound has peak areas resolvable from noise.

$$A_{16} = \sum_{i=1}^{5} {}^{1}A_i, \quad A_{18} = \sum_{i=1}^{5} {}^{2}A_i + \sum_{i=1}^{5} {}^{3}A_i$$

(Equation 2)

The equations 3 calculate the peak areas for unlabeled and ^{18}O labeled peptides, considering each single isotopic peak as a fraction of the first isotopic peak. The theoretic isotopic distributions for all peptides used in quantification experiments were calculated using program PeptideProspector 4.0.4 (University of California, http://prospector.ucsf.edu), presuming that the differences in isotopic distribution ratios for ^{18}O-labeled and unlabeled peptide are negligible.

$$A_{16} = \sum_{i=1}^{5} f_i \cdot {}^{1}A_i, \quad A_{18} = \sum_{i=1}^{5} f_i \cdot {}^{2}A_1 + \sum_{i=1}^{5} f_i \cdot {}^{3}A_1$$

(Equation 3)

The obtained extended equation 1 can be further rearranged and simplified as it can be seen in equation 4:

$$n_{16} = \frac{\sum_{i=1}^{5} f_i \cdot {}^{1}A_1}{\sum_{i=1}^{5} f_i \cdot {}^{2}A_1 + \sum_{i=1}^{5} f_i \cdot {}^{3}A_1} \cdot n_{18} \Rightarrow n_{16} = \frac{{}^{1}A_1}{{}^{2}A_1 + {}^{3}A_1} \cdot n_{18}$$

(Equation 4)

The equations 5 and 6 express the peak areas for the first isotopic peaks of the single and double labeled forms of the peptide ($^{*2}A_1$ and $^{*3}A_1$, peak areas of the peaks m/z 1569.939 and m/z 1571.938 in Figure 2.11 a):

$${}^{*2}A_1 = {}^{2}A_1 + {}^{1}A_3 = {}^{2}A_1 + f_3 \cdot {}^{1}A_1 \Rightarrow {}^{2}A_1 = {}^{*2}A_1 - f_3 \cdot {}^{1}A_1$$

(Equation 5)

$$\begin{aligned} &{}^{*3}A_1 = {}^{3}A_1 + {}^{2}A_3 + {}^{1}A_5 = {}^{3}A_1 + f_3 \cdot {}^{2}A_1 + f_5 \cdot {}^{1}A_1 = {}^{3}A_1 + f_3 \cdot ({}^{*2}A_1 - f_3 \cdot {}^{1}A_1) + f_5 \cdot {}^{1}A_1 \Rightarrow \\ &\Rightarrow {}^{3}A_1 = {}^{*3}A_1 - f_3 \cdot {}^{*2}A_3 - {}^{1}A_1 \cdot (f_5 - f_3^2) \end{aligned}$$

(Equation 6)

Finally, establishing these expressions into the equation 4, we obtain after rearrangement equation 7, which can be used for quantification calculations:

$$n_{16} = \frac{{}^{1}A_1}{{}^{*3}A_1 + {}^{*2}A_1 \cdot (1 - f_3) + {}^{1}A_1 \cdot (f_3^2 - f_5 - f_3)} \cdot n_{18}$$

(Equation 7)

In this equation ${}^{*3}A_l$ and ${}^{*2}A_l$ are the areas of the isotopic peaks spaced from the monoisotopic peak of the unlabeled peptide by 2 and 4 Da, respectively (peaks 1569.94 and 1571.94 in Figure 2.11 a). Coefficients f_3 and f_5 are the calculated ratios of the intensity of, respectively, third (+ 2 Da) and fifth (+ 4 Da) isotopic peaks to the intensity of the monoisotopic peak of the unlabeled peptide ($f_3 = 0.46$ and $f_5 = 0.05$ for the peptide DAFLGSFLYEYSR (m/z 1567.74) in Figure 2.11).

The above described calculations take into account both the incomplete incorporation of ^{18}O and differences in natural isotope distributions of individual peptides.

2.1.3.2 What factors affect labeling stability and efficiency?

Schnolzer et al. [262] systematically studied enzyme-catalyzed ^{18}O-labeling of peptides during proteolytic digestion and reported that trypsin, Lys-C, and Glu-C incorporate two ^{18}O atoms into the carboxyl termini of all peptides, except the original protein carboxyl termini. On the contrary, some reports indicate that lysine-terminated peptides do not incorporate two oxygen labels efficiently [295, 296].

Application of ^{18}O-labeled internal standards requires their stability and therefore their general exchange characteristics should be well understood in order to ensure analytical accuracy of protein quantification. Therefore I investigated some parameters, which affect the labeling efficiency, including effect of pH and nature of the peptide. Labeling efficiency (0-100%) in this context refers to the degree to which a peptide is labeled with one or two ^{18}O atoms. Labeling can be considered 100% if there remains no unlabeled peptides, i.e. at least one ^{18}O atom is incorporated. Further, the efficiency can be then differentiated between singly and doubly labeled peptides, whereas 100% double labeling is the maximum labeled state.

The effect of pH on the stability of ^{18}O-labeled peptide was studied by digestion of BSA in 95% atom abundance $H_2{}^{18}O$ and subsequent dilution with ^{16}O water containing formic acid (FA), so that its amount in the mixture ranged from 0 to 5% (v/v). Formic acid is used to reduce the pH and stop the digestion by inhibiting trypsin activity. Consistent with previously reports my results [262, 294] confirmed that under pH higher than 5 (low content of FA, < 1 % (v/v)) the enzyme is still active and continues to catalyze the back-exchange in medium containing ^{16}O. To avoid back-exchange by mixing of ^{18}O-labeled internal standard with the sample, containing ^{16}O acidic conditions (pH 3-4) should be maintained.

Then I studied the relative labeling efficiency (labeled to non-labeled) and relative degree of labeling ($^{18}O_1/^{18}O_2$) for different peptides generated by digestion of BSA in 95% $H_2{}^{18}O$. To this end, I monitored the changes in ^{18}O labeling in dependence on digestion time (30 min, 1.5h, 3h and overnight digestion) and temperature (digestion temperatures were 22, 37 and 55°C). The labeling efficiency ($^{16}O/^{18}O$) was in all cases close to 95% as is consistent with enzyme-catalyzed hydrolysis of protein. However, the degree of labeling ($^{18}O_1/^{18}O_2$) was less consistent. My experiments showed that incorporation of second ^{18}O was less pronounced by peptides containing lysine-termini (Figure 2.12). Although, the degree of labeling with two ^{18}O of arginine terminated peptides was in general close to 92% (at all tested temperatures and hydrolysis times; Figure 2.13, a), peptides terminated by lysine showed in major cases slow incorporation of the second ^{18}O into their carboxyl termini and strong dependence on the digestion temperature (Figure 2.13, b). While lysine containing peptide FKDLGEEHFK (M_R=1248.61) incorporated about 30% of ^{18}O in double-labeled form upon 1.5h digestion at temperature 22 and 37°C and about 80% upon overnight digestion at the same temperatures, at higher temperature (55°C) the labeling degree with $^{18}O_2$ increased from 5 to 10% during 1.5h of digestion and remained constant, as expected, due to the thermal deactivation of trypsin.

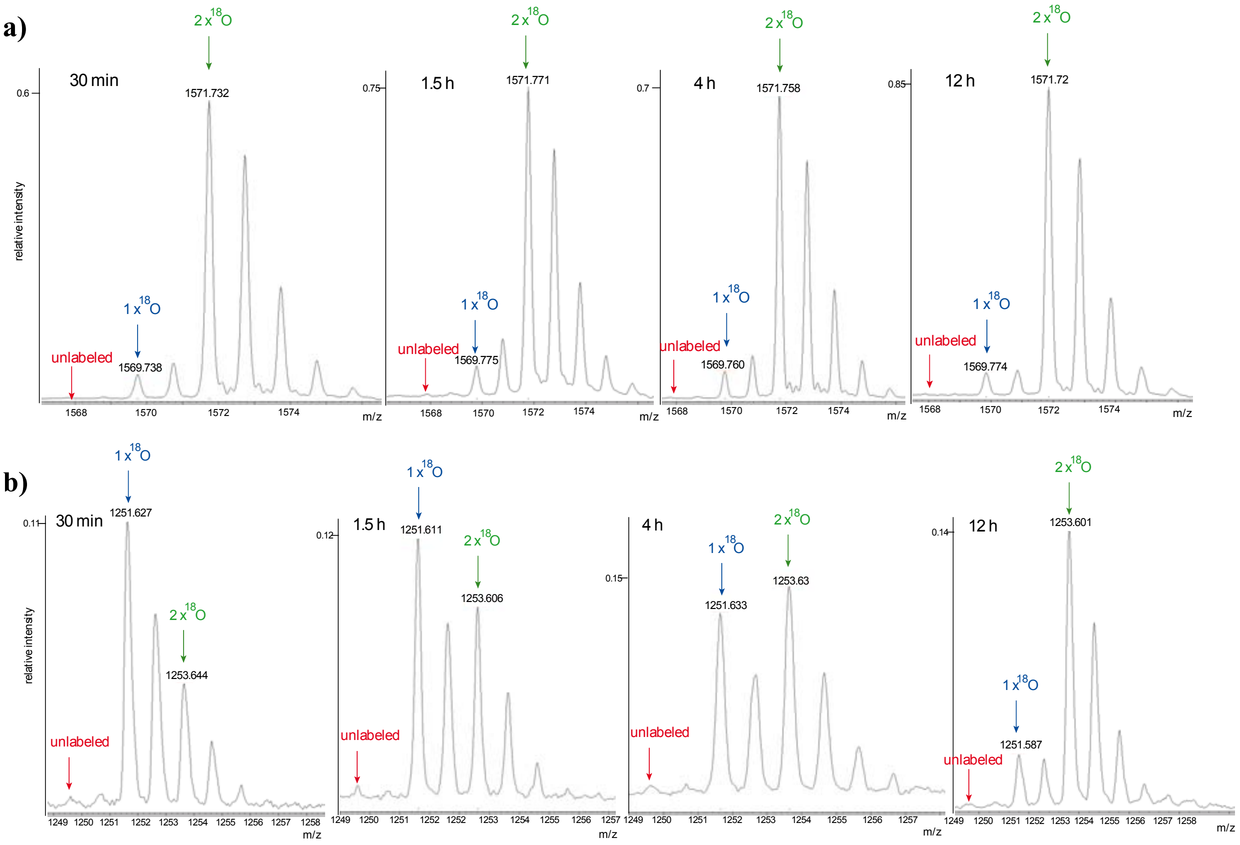


Figure 2.12: MALDI TOF spectra of peptides DAFLGSFLYEYSR (m/z 1567.74) and FKDLGEEHFK (m/z 1249.61) obtained by BSA tryptic digestion in the buffer containing 95% $H_2{}^{18}O$.

(a) and (b) MALDI TOF spectra of peptides DAFLGSFLYEYSR (m/z 1567.74) and FKDLGEEHFK (m/z 1249.61), respectively; digestion was performed at 37°C and terminated with 5% formic acid after 0.5; 1.5; 4 and 12 h of incubation. Arrows demonstrate monoisotopic peaks of unlabeled (red), single-^{18}O-labeled (blue) and double-^{18}O-labeled (green) peptides.

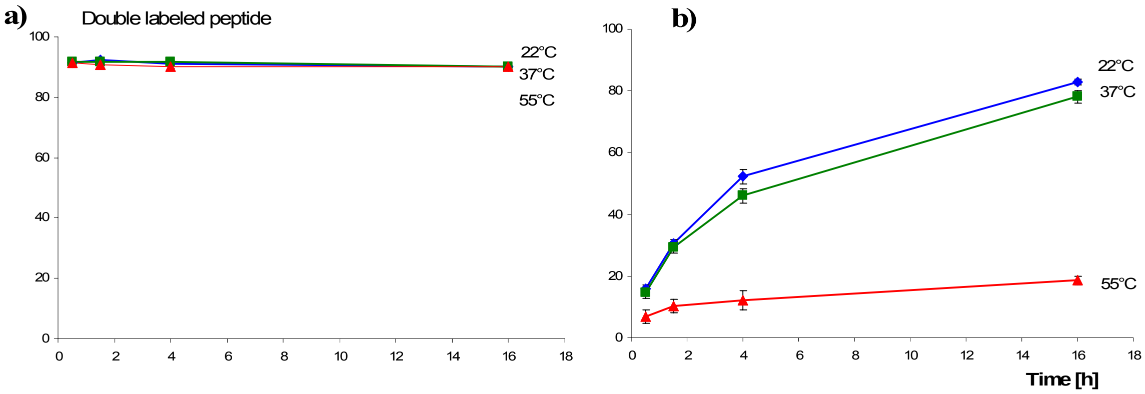


Figure 2.13: Degree of labeling ($^{18}O_2$ / $^{18}O_1$) for peptides DAFLGSFLYEYSR (m/z 1567.74) and FKDLGEEHFK (m/z 1249.61) obtained by BSA tryptic digestion in the buffer containing 95% $H_2{}^{18}O$ at different temperatures.

(c) and (d) degree of labeling ($^{18}O_2$ / $^{18}O_1$) for peptides DAFLGSFLYEYSR (m/z 1567.74) and FKDLGEEHFK (m/z 1249.61), respectively; coloured curves represent $^{18}O_2$ / $^{18}O_1$ obtained for both peptides at different digestion temperatures (red 55°C, green 37°C and blue 22°C).

My results were consistent with several researches, which showed that the degree of ^{18}O labeling is not universal consistent from peptide to peptide and is dependent on the nature of the peptide [294-296]. Schnolzer et al. [262] suggested the mechanism of enzyme-catalyzed ^{18}O labeling. This process can be divided into two parts: 1) cleavage of the peptide amid bond by formation of acyl-enzyme intermediate at the C-terminus of the newly formed peptide, which is then hydrolyzed to form the free peptide and 2) incorporation of the second ^{18}O atom by reformation of a peptide-trypsin ester complex and its subsequent hydrolysis [262] (Figure 2.14). The incorporation of a second ^{18}O atom is dependent, whether a formed peptide fragment is accepted as a pseudo-substrate ester intermediate and is dependent on the nature of peptide. My results suggest that lysine-terminated peptides have poorer enzyme-substrate selectivity compared to peptides terminated by arginine, what results in less efficient incorporation of a second ^{18}O.

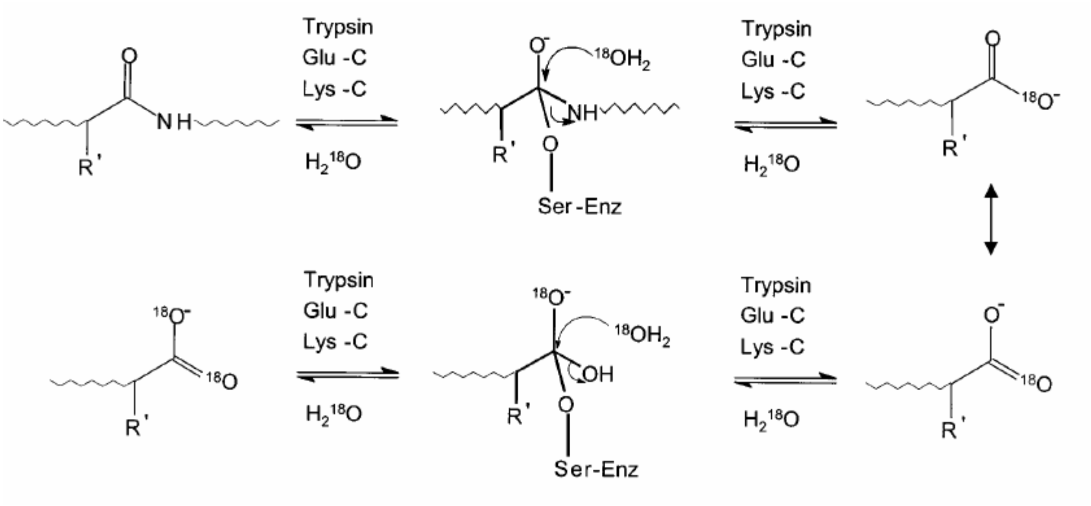


Figure 2.14: Mechnism of enzyme-catalyzed ^{18}O incorporation during proteolysis.

(adapted from Yao, X et al.) Upper part includes formation of acyl-enzyme intermediate at the C-terminus of the newly formed peptide, which is then hydrolyzed to form the free peptide. Lower part includes the incorporation of the second ^{18}O atom: a peptide-trypsin ester complex is reformed and subsequently hydrolyzed.

Since arginine terminated peptides show more than 90 % double labeling they can be used for absolute quantification without complex deconvolution method. So the equation 7 can be simplified to equation 8:

$$n_{16} = \frac{{}^{1}A_1}{{}^{*3}A_1} \cdot n_{18}$$

(Equation 8)

In addition, MALDI peptide mass fingerprints of tryptic digests are dominated by peptides containing arginine residues at the C-terminus, due to differential ionization effects and the basicities of arginine- and lysine-containing peptides [25, 297]. Therefore for my kinetics studies following three Arg-containing peptides were applied: YLYEIAR, m/z 927.49; LGEYGFQNALIVR, m/z 1479.79; and DAFLGSFLYEYSR m/z 1567.74 (Figure 2.15). The yield of digestion was calculated by averaging the amount of these peptides.

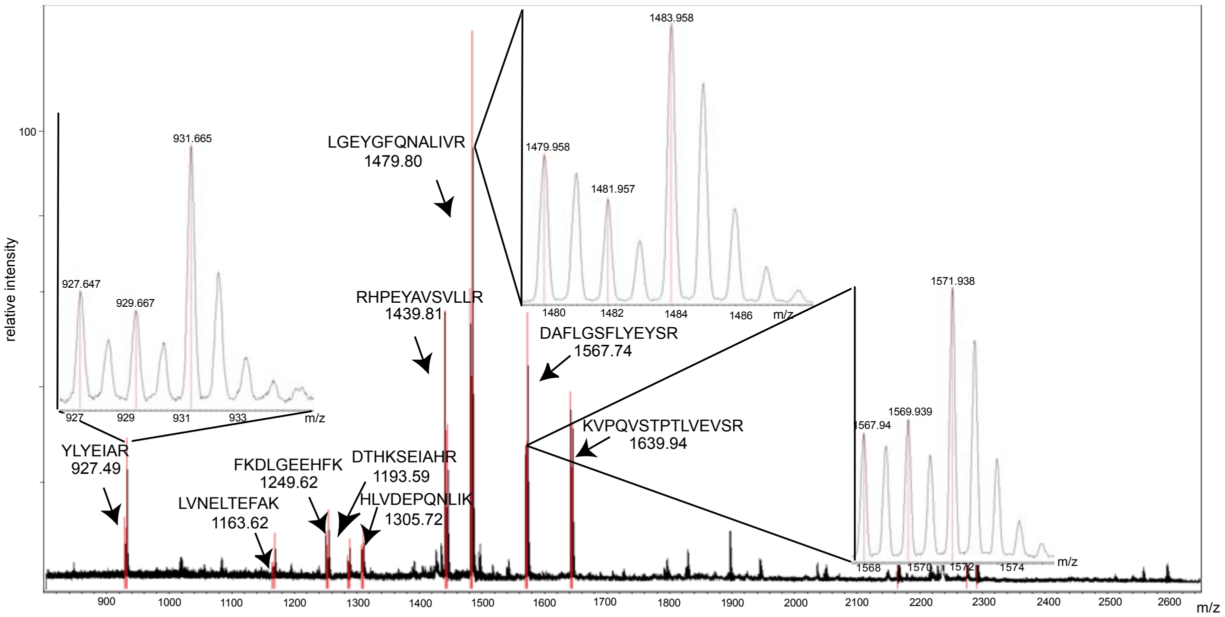


Figure 2.15: MALDI TOF spectrum of the mixture containing BSA peptides obtained by in-gel digestion and their ^{18}O-labeled internal standards.

BSA (5pmol) was in-gel digested; the tryptic peptides were extracted and the extract was dried down. The obtained peptides were redissolved in 10 µL of 0.3 µM ISTD prepared by BSA tryptic digestion in the buffer containing 95% $H_2{}^{18}O$. Tryptic peptides of BSA are depicted in the spectrum by arrows with corresponding peptide sequences and m/z calculated for the unlabeled monoisotopic ions. Blowouts demonstrate merged isotopic clusters of the peptides YLYEIAR, m/z 927.49; LGEYGFQNALIVR, m/z 1479.79; DAFLGSFLYEYSR m/z 1567.74 and their ^{18}O-labeled internal standards.

2.1.3.3 ^{18}O labeling approach: what is important?

^{18}O labeling approach investigated above provides a relatively simple and sensitive method for absolute quantification of proteins in a variety of proteomic applications. It has been demonstrated that under mild acidic conditions typically used for ESI- and MALDI-MS, ^{18}O-containing carboxyl groups of peptides are sufficiently stable. Theoretically, all labeled tryptic peptides can be applied for quantification. However, in practice, the number of peptides which can be used for quantification by MALDI TOF MS is lower, due to lower (more than 5 times) signal intensities of lysine- (compared to Arg) containing peptides and their poorer enzyme-substrate selectivity, which slows incorporation of second ^{18}O into the C-terminal carboxylic acid (so requiring deconvolution of isotopic clusters). For this reasons, arginine-containing peptides should be favoured in proteomic quantification studies.

2.1.3.4 Kinetic study: effect of digestion time and enzyme concentration on the recovery of tryptic peptides

In the following kinetic study I first monitored the time course of in-gel digestion at elevated temperature. To this end BSA gel bands (5pmol) were digested for 0.5h, 1.5 h, 3 h, and overnight at 55°C by glycosylated trypsins (Figure 2.16); concentration of the modified enzymes was on average 0.5 µM and the yield of conventional digestion (37 °C, overnight, by native BT at the concentration ~ 0.5 µM) was used as a reference. Between 5 and 20% of conventional digestion yield was reached in 30 min of digestion by RAF-BT, MAT-BT and STA-BT; about 30 % was achieved by RAFR-BT. Overnight cleavage resulted about 40 % of conventional digestion yield for STA-BT and between 50 and 70 % for MAT-BT, RAF-BT and RAFR-BT (Figure 2.16).

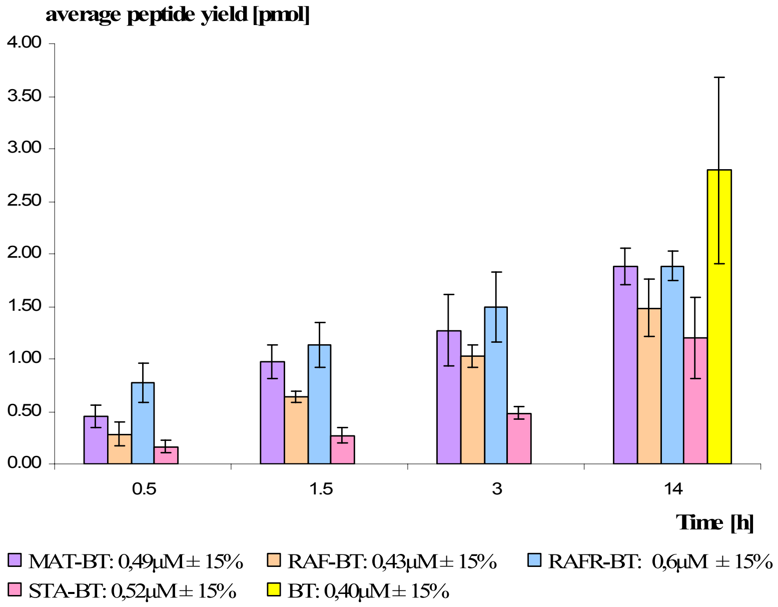


Figure 2.16: Time course of averaged peptide yield observed by in-gel digestion of BSA at elevated temperature using glycosylated trypsins at an enzyme concentration in average 0.5 µM.

BSA bands (5 pmol) were in-gel digested by glycosylated trypsins at 55°C; digestion times were: 30, 1.5h, 3h, and 12h. The applied enzyme concentrations ranged from 0.43µM for RAF-BT to 0.6µM for RAFR-BT (enzyme stock solutions were subjected to amino acid analysis in the laboratory of Dr. Hunziker (University of Zürich, Switzerland)). The coloured bars represent the digestion yields generated by trypsin conjugates (purple, pale pink, blue and pink); yellow bar represents the recovery of the conventional digestion (37°C, overnight, BT at concentration ~ 0.5 µM).

According to the enhanced in-gel digestion protocol developed by Havliš et al. [2] I set out to investigate whether the low yield of 30 min digestion can be improved by increasing the enzyme concentration. T_{50} constants of the glycosylated trypsins (defined as a temperature at which 50% of the enzyme activity is retained upon 30 min incubation) are by ~20°C higher than T_{50} constant of unmodified BT. I assumed that the partial deactivation and autolysis of the enzymes for a short time might not negatively impact the yield of digestion and might not overpopulate spectrum with autolytic products. BSA gel bands were digested for 30 minutes at different enzyme concentrations ranging from 0.5 to 3 μM. Consistently with previously reported results [2] about 65 % of the yield of conventional digestion was reached by protein digestion with methylated porcine trypsin (Promega) at enzyme concentration of 0.75 μM and more than 100 % at concentration between 1 and 1.5μM. Protein digestion by MAT-BT, RAF-BT and STA-BT at concentrations between 0.5 and 1 μM resulted in less than 20% of conventional digestion yield and achieved for RAFR-BT about 40% (Table 2.3). As anticipated, the digestion yield increased with increased concentrations of modified enzymes. The recovery of 40 to 50% of conventional digestion was reached by all glycosylated trypsins at enzyme concentrations between 2 and 3 μM. However, the digests were strongly contaminated with trypsin autolysis products, which complicated protein identification by MALDI MS.

Figure 2.17 presents a peptide mass fingerprint of a BSA in-gel digest (5 pmol) obtained by accelerated digestion using RAF-BT at the highest tested enzyme concentration (2.2 μM). Although BSA amount was relatively high peptide mass fingerprint of analyzed digest contained abundant autolytic trypsin peptides LGEDNIN-VVEGNEQFISASK (m/z 2163.1), SIVHPSYNSNTLNNDIMLIK (m/z 2273.2) and SIVHPSYNSNTLNNDIM(ox)LIK (m/z 2289.2), which overloaded the spectrum and complicated peak picking. Thus, many of detected tryptic peptides of BSA were at the noise level. Most intense BSA peaks were obtained from peptides YLYEIAR (m/z 927.49), RHPEYAVSVLLR (m/z 1439.8), LGEYGFQNALIVR (m/z 1479.79), DAFLGSFLYEYSR (m/z 1567.74) and KVPQVSTPTLVEVSR (m/z 1639.93). However, their intensity was significantly lower than the intensity of the autolytic background of trypsin. Altogether 10 BSA peptides were identified by MASCOT database searching, covering 20% of the protein sequence.

I found it was impossible to apply trypsin conjugates at concentrations higher than 1.5 µM. The enzyme concentration of about 1 µM was considered as optimal concentration for accelerated in-gel digestion by glycosylated trypsin conjugates.

Table 2.3: Averaged peptide yield of BSA obtained upon 30 min of accelerated in-gel digestion by glycosylated trypsins at different enzyme concentrations.

Modified enzyme	Concentration [µM] [a]	Yield [b] [pmol]	SD [c]	RSD [d]	Recovery [%] [e]
native BT [f]	0.5	2.80	0.89	32	
MET-PT	0.75	1.79	0.78	43	64
	1.5	3.33	0.83	25	119
MAT-BT	0.5	0.45	0.11	24	16
	1.0	0.37	0.08	21	13
	2.5	1.22	0.37	30	44
RAF-BT	0.4	0.29	0.12	42	10
	0.9	0.54	0.08	15	19
	2.2	1.09	0.16	15	39
RAFR-BT	0.6	0.77	0.19	24	28
	1.2	1.00	0.19	19	36
	3.0	1.47	0.42	28	53
STA-BT	0.5	0.17	0.06	36	6
	1.0	0.36	0.14	39	13
	2.5	1.20	0.20	17	43

[a)] Concentration determined by amino acid analysis;
[b)] Average peptide yield of the digestion;
[c)] Standard deviation of the calculated peptide yield; [d)] Relative standard deviation of the calculated peptide yield;
[e)] Percentage of the tryptic peptide recovery of conventional digestion;
[f)] Conventional digestion (overnight, at 37 °C and at concentration of BT ~ 0.5 µM).

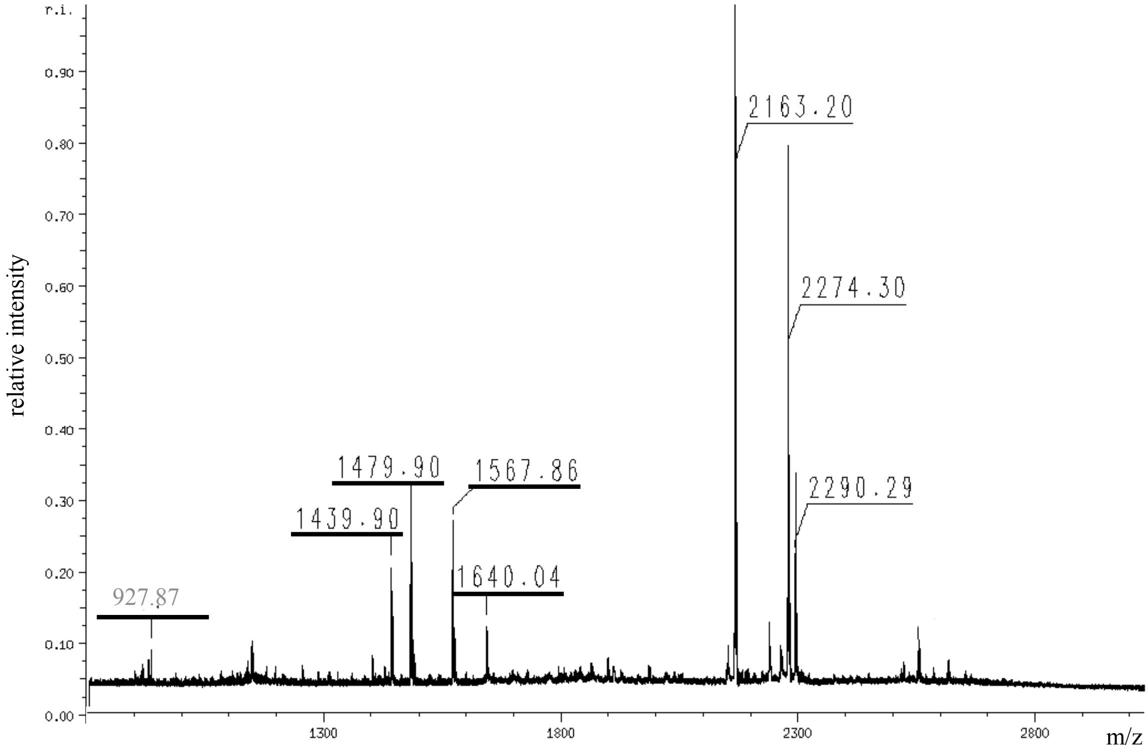


Figure 2.17. Peptide mass fingerprint of BSA in-gel digest obtained by accelerated digestion using RAF-BT at high concentration.

A gel band containing 5 pmol BSA was digested by RAF-BT; the digestion time was set at 0.5h and the applied enzyme concentration was 2.2 µM. Bold underlined peaks represent most intense BSA peptides: YLYEIAR, m/z 927.49; RHPEYAVSVLLR, m/z 1439.8; LGEYGFQNALIVR, m/z 1479.79; DAFLGSFLYEYSR, m/z 1567.74; KVPQVSTPTL-VEVSR, m/z 1639.93, which were identified by MASCOT search. High abundant peaks represent autolytic peptides of trypsin: LGEDNINVVEGNEQFISASK (m/z 2163.1), SIVHPSYNSNTLNNDIMLIK (m/z 2273.2) and SIVHPSYNSNTLNNDIM(ox)LIK (m/z 2289.2).

As next, I monitored the time course of in-gel digestion using trypsin conjugates at an enzyme concentration of in average 1 µM (Figure 2.18). My intension was to investigate whether longer digestion times may improve the digestion yield without compromising the quality of the spectrum due to increased autolysis background.

The peptide recovery after 1.5 h of digestion achieved 20% for STA-BT and 60% for RAFR-BT of conventional digestion yield, digestion for 3h only gained about 10% of increase. Considering that the autolysis products of the trypsin conjugates under these conditions didn’t overload the spectrum and were not disturbing for MALDI TOF analysis the optimal digestion time can be set at 3 hour.

Table 2.4 represents the peptide yield of in-gel digestion (5 pmol BSA) performed by glycosylated trypsins after 3 hours of incubation at 55°C. From 30% (for STA-BT) to 65% (for RAFR-BT) of conventional digestion yield could be reached by accelerated digestion for 3 hours at an enzyme concentration about 1 µM. In contrast to enhanced in-gel digestion method, which applies methylated porcine trypsin in-gel digestion by glycosylated trypsin conjugates didn’t reach the yield of the conventional digestion.

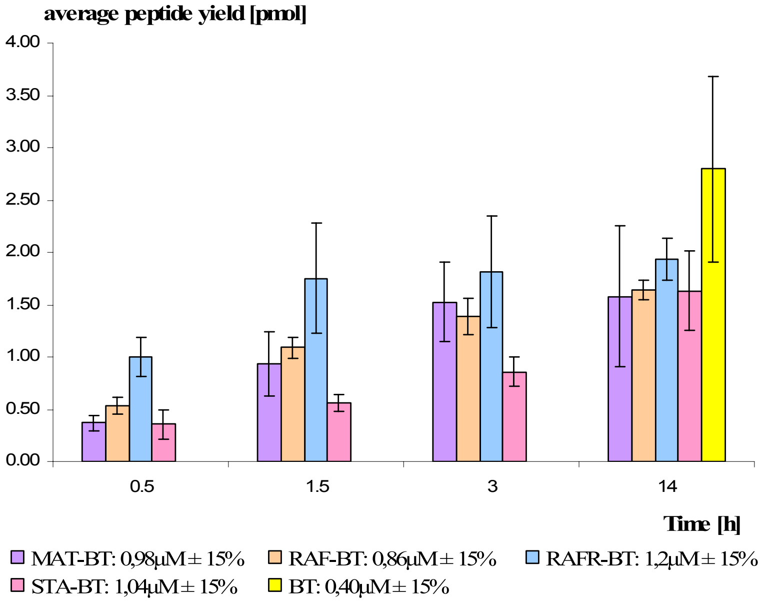


Figure 2.18: Time course of averaged peptide yield observed by in-gel digestion of BSA using glycosylated trypsins at elevated temperature and an enzyme concentration in average 1µM.

BSA bands (5 pmol) were in-gel digested by glycosylated trypsins; digestion times were: 30 min, 1.5 h., 3 h., and 12 h. The applied enzyme concentrations ranged from 0.86µM for RAF-BT to 1.2µM for RAFR-BT. The coloured bars represent the digestion yields generated by trypsin conjugates (purple, pale pink, blue and pink); yellow bar represents the recovery of the conventional digestion (37°C, overnight, BT at concentration ~ 0.5 µM).

Table 2.4: : Averaged peptide yield of BSA obtained upon 3 hours of accelerated in-gel digestion by glycosylyted trypsins at enzyme concentration ~1µM.

Modified enzyme	Concentration [µM] [a]	Yield [b] [pmol]	SD [c]	RSD [d]	Recovery [%] [e]
native BT [f]	0.5	2.80	0.89	31.67	
MAT-BT	1	1.53	0.37	24.50	55
RAF-BT	0.9	1.38	0.17	12.58	49
RAFR-BT	1.2	1.81	0.53	29.38	65
STA-BT	1	0.86	0.14	16.32	31

[a)] Concentration determined by amino acid analysis; [b)] Average peptide yield of the digestion; [c)] Standard deviation of the calculated peptide yield; [d)] Relative standard deviation of the calculated peptide yield; [e)] Percentage of the tryptic peptide recovery of conventional digestion; [f)] Conventional digestion (overnight, at 37 °C and concentration of BT ~ 0.5 µM).

2.1.3.5 Effect of gel pore size on the yield of in-gel digestion

Since modification of bovine trypsin with saccharides increased its molecular weight and consequently changed its diffusion mobility I set out as next to investigate, how the gel pore size influences the yield of in-gel digestion by glycosylated trypsin conjugates. To this end I studied kinetics of digestion in 8 and 12 % polyacrylamide gel matrix. Figure 2.19 shows digestion yield of BSA peptides generated in 8 and 12 % polyacrylamide gels after 30 min (Figure 2.19A) and 3 h (Figure 2.19B) incubation with RAFR-BT. As shown in the Figure 2.19 no changes have been obtained in the digestion yields generated in 8 and 12% polyacrylamide matrixes.

Average pore size ranges for 8% polyacrylamide gels from 16 to 22 Å and for 12% polyacrylamide gels from 10 to 15 Å [298-300]. The molecule size of the most bulky trypsin modification with ACD (M_R of ACD-BT determined by tricine-SDS-PAGE was ~32 kDa) was close to 20 Å [291]. Trypsin conjugates modified with ACD and BCD were, as expected, not efficient in digestion in 8% as well as in 12% polyacrylamide gels. Tested RAFR modified trypsin (M_R of RAFR-BT determined by SDS-PAGE was ~28 kDa) has smaller molecule size compared to bulky ACD and BCD-BT. However, considering quite broad range of pore size in 8 and 12 % polyacrylamide gels, it is difficult to predict whether 8% gels allow unrestricted diffusion of conjugated enzyme molecules into the gel matrix.

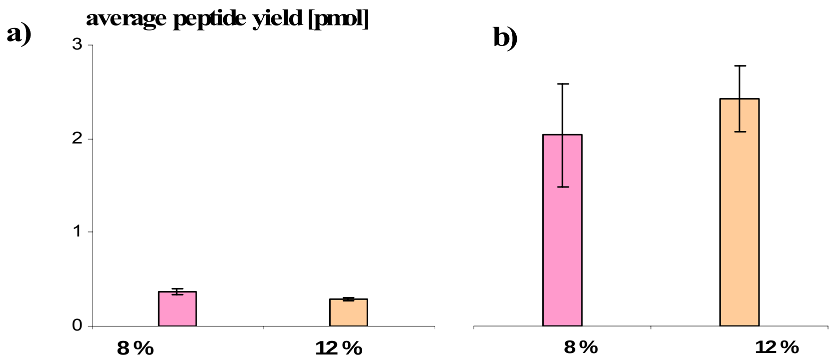


Figure 2.19: Effect of polyacrylamide gel pore size on the digestion yield.

BSA bands (5 pmol) were in-gel digested by RAFR-BT at enzyme concentration of 1.2µM; digestion time was set at (A) 30 min and (B) 3 hours. Coloured bars represent the digestion yield: pink in 8% polyacrylamide gel; pale pink in 12% polyacrylamide gel.

The obtained results rather suggest that the molecule size of the glycosylated trypsin conjugates is of borderline value and that their bulky structure limits their diffusion mobility in in-gel digestion of proteins.

2.1.3.6 Proof of the method

I then applied RAF modified trypsin in an ongoing collaborating project under accelerated digestion conditions. The efficiency of accelerated digestion was directly compared with the efficiency of conventional digestion method by identification of members of a NuA4 histone acetyltransferase complex isolated from the budding yeast. A protein YNG2 was epitope-tagged and immunoaffinity purified using the tandem affinity purification (TAP) method (performed by Luke Buchanan from Prof. Francis Stewart laboratory, BIOTEC, Dresden). The purified protein complex was separated by one-dimensional gel electrophoresis and the protein bands were visualized by Coomassie staining. 11 protein bands in the range of 20-140 kDa, with protein content of 0.5 – 2 pmol (according to the staining intensity of BSA standards), were excised from the gel and each band was cut into two parts. One part of the protein band was digested by RAF-BT at 55 °C for 3 h at an enzyme concentration of ~1µM and another part was digested using conventional method by unmodified trypsin (overnight digestion at 37 °C and an enzyme concentration of BT ~ 0.5 µM). The samples were subsequently analyzed by MALDI TOF mass spectrometry. To evaluate the quality of the spectra and the significance of protein identification I used MOWSE scores (statistical significance measure for protein identification provided by MASCOT database search program), sequence coverage of identified proteins, number of peaks manually picked from the spectrum, and number of peaks matched to the sequence of identified proteins (Table 2.5).

Although all proteins were successfully identified by both methods, conventional digestion by unmodified trypsin showed better performance than accelerated digestion by raffinose-modified trypsin. Sequence coverage of proteins identified by fast digestion method was lower than by conventional digestion. On the other hand, RAF-BT showed less autolytic background in the mass fingerprints and so simplified the identification of low abundant proteins that contained few peptides.

Table 2.5: Comparison of conventional digestion by unmodified bovine trypsin and accelerated digestion by RAF-BT.

			Conventional digestion by unmodified BT				Accelerated digestion by RAF-BT			
Band	Protein	MW	peaks	peaks	MASCOT	Sequence	peaks	peaks	MASCOT	Sequence
		kDa	total[a]	matched[b]	score[c]	coverage,%	total[a]	matched[b]	score[c]	coverage,%
1	TRA1	432	113	73	241	20	85	47	87	13
2	VID21	112	112	58	363	51	102	57	335	52
3	EPL1	97	121	38	192	47	97	32	144	39
4	SSA2	69	93	22	106	36	76	18	75	33
5	SSB1	66	85	22	117	35	70	17	86	41
6	SWC4	55	105	26	146	42	86	18	95	30
7	ESA1	53	93	20	114	36	83	20	111	39
8	EAF3	45	83	20	119	41	66	20	119	41
9	ACT1	41	76	14	75	42	51	15	75	36
10	EAF5	32	100	19	121	66	87	15	88	50
11	YAF9	26	83	14	108	65	85	16	102	59

Applied digestion conditions: conventional digestion was performed overnight at 37°C by commercially available BT (Roche) at an enzyme concentration of ~ 0.5 µM; accelerated digestion by RAF-BT was performed for 3 h at 55°C and at an enzyme concentration of ~ 1 µM.

[a)] Total number of peaks picked; [b)] number of peaks matched the sequence of identified proteins; [c)] MOWSE score.

2.1.3.7 Catalytic efficiency of trypsin conjugates in accelerated in-gel digestion: what did we learn?

The described kinetic study provided evaluation of the effect of digestion conditions on the yield of in-gel digestion performed using glycosylated trypsin conjugates. This study enabled adjustment of the optimal reaction conditions as in ACD previously established by Havliš et al. [2].

Although glycosylated trypsins showed better thermostability compared to conventional unmodified bovine trypsin and methylated porcine trypsin, their catalytic efficiency in in-gel digestion was lower. As expected, the efficiency of modified trypsin derivates in in-gel digestion strongly depends on their modification. Enzymes caring larger oligosaccharides, such as a tetrasaccharide stachyose and cyclodextrins are more rigid and bulky, and consequently have lower diffusion mobility than enzymes modified with disaccharides maltotriose and raffinose, resulting in lower digestion yield. On the other hand, poor yield of in-gel digestion by saccharide modified trypsins can be explained by their lower specific activity (by 10-30%) compared to those of unmodified trypsin.

2.1.4 Label-free quantification by nanoLC-MS/MS

The quantitative study of digestion kinetics described above is based on stable isotope labeling of peptides with ^{18}O (used as internal standards) and MALDI TOF MS analysis. This approach, however, has some limitations. First, it is expensive. Second, the number of peptides, which can be used for quantification, is rather low because the absolute intensities of the detected ions depend on their chemical nature and suppression effects occur. Spectral quality is also greatly affected by the method of sample preparation, MALDI matrix composition, and possible sample impurities.

Label-free protein quantification methods offer less expensive and simple sample handling. Several studies have demonstrated that mass spectral peak intensities of peptide ions obtained from LC-MS/MS data correlate well with protein abundances in complex samples [272-276]. Therefore I set out to investigate the performance of this quantification method in nanoLC-MS/MS analysis. Further I aimed to test this approach in the study of digestion kinetics as it was previously carried out by MALDI TOF MS using ^{18}O-labeled peptides as internal standards.

2.1.4.1 Quantifying proteins by mass spectrometric signal intensities of their peptide ions

a) Study of a single protein

To check the performance of this quantification method in nanoLC-MS/MS analysis I first started with study of a single protein. BSA was enzymatically digested and 8 aliquots of its serial dilutions containing protein amounts from 6 to 750 fmol were separated on a 75 µm i.d. reversed-phase column and directly electrosprayed (via a dynamic nanospray probe) into a LTQ ion trap mass spectrometer (Thermo Electron Corp.), which was operated in data-dependent acquisition mode (see chapter 4.1.3.4). The experiment was repeated on five consecutive days in order to evaluate the analytical reproducibility of MS signal and retention time.

To calculate peptide ion intensities extracted ion chromatograms (XICs) were generated from the full scan mass spectra within a narrow m/z range, corresponding to different charge states of a peptide (triple, double, and single). The ion intensity of a peptide was subsequently calculated by summing peak areas of its triple, double, and single charged ions.

Narrow m/z range has to be chosen in order to minimize number of peaks in XICs generated by peptides, which have similar m/z values. On the other hand, the mass tolerance for precursor ions (given for the applied MS instrument) should be considered and selected m/z range should enable inclusion of at least 3 isotopic peptide peaks. To this end I used m/z range with

a lower limit of ***m/z = ((peptide monoisotopic mass – 1) + charge)/charge*** and

upper limit of ***m/z = ((peptide monoisotopic mass + 2) + charge)/charge***.

The correct retention time of a peptide was determined from the scan number of its MS/MS spectra confidently identified by MASCOT search.

Figure 2.20 a shows base peak ion chromatogram of a BSA tryptic digest and XICs of its peptide LVDEPQNLIK eluted at 32.23 min, which was represented by triple, double, and single charged states (Figure 2.20 b, c and d). Peptides characterized in this analysis are depicted in the table 2.6 among with their retention times, observed charge states and calculated peak areas.

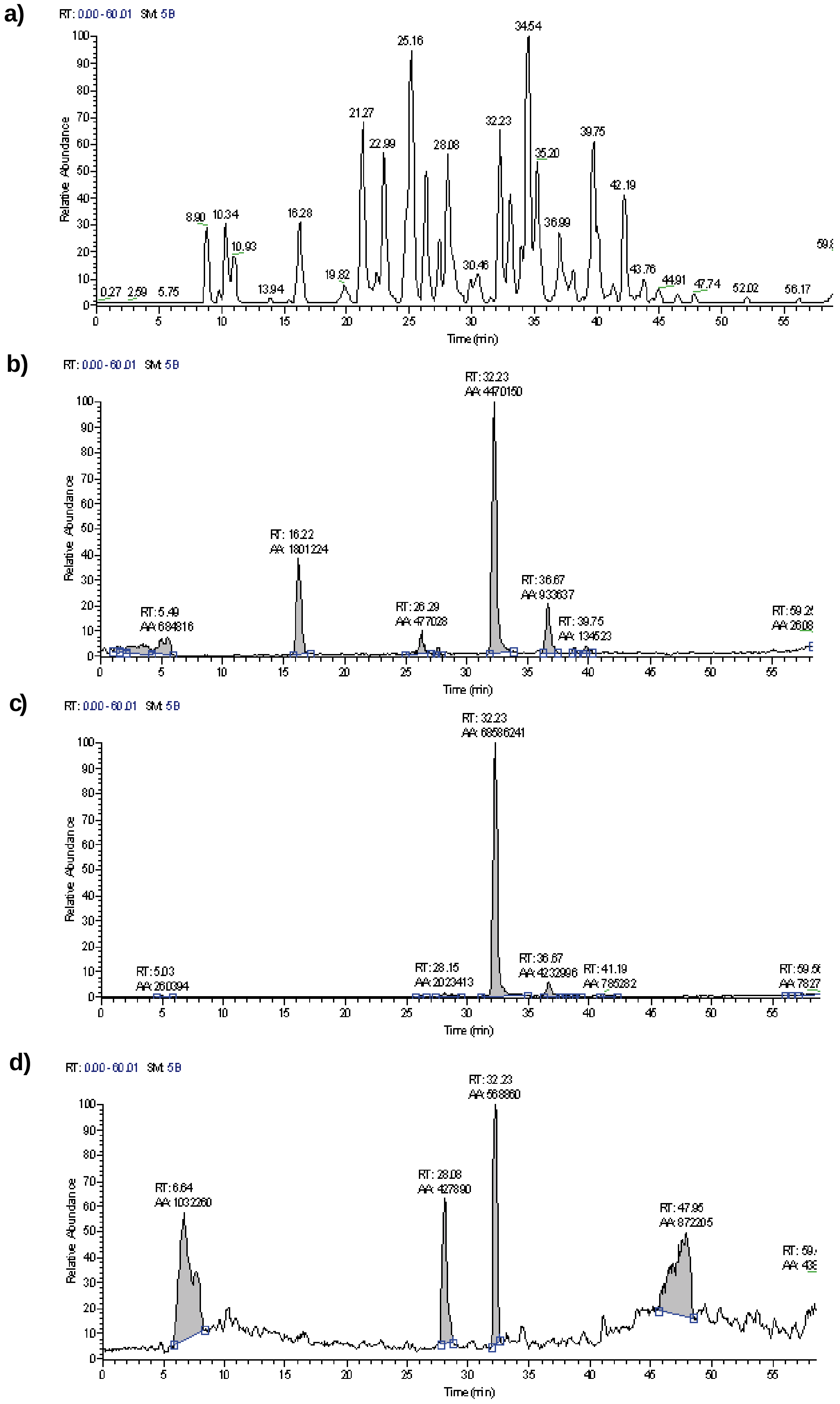


Figure 2.20: Base peak ion chromatogram of a BSA tryptic digest and extracted ion chromatograms of differently charhed ions of BSA peptide LVDEPQNLIK.

a) Base peak ion chromatogram of a BSA tryptic digest; the amount of protein analyzed by nanoLC-MS/MS was 188 fmol. (b), (c), (d) XICs of the triple, double and single charged ions of BSA peptide HLVDEPQNLIK, respectively. Scan numbers corresponding to the MS/MS spectra of this peptide helped to identify its correct retention time, at 32.23 min.

Table 2.6: Peptides characterized in nanoLC-MS/MS analysis of the BSA tryptic digest, including m/z values and corresponding charge states, calculated peak areas and retention times.

no.	Peptide	Molecular weight	Charge	m/z	Peak area	Retention time
1	AWSVAR	688.37	2+ 1+	345.19 689.37	2.18E+07 3.92E+06 2.57E+07	24.65
2	GACLLPK	757.42	2+ 1+	379.71 758.42	1.18E+07 2.07E+06 1.39E+07	26.49
3	LVTDLTK	788.46	2+ 1+	395.23 789.46	2.22E+07 5.08E+06 2.73E+07	24.78
4	LSQKFPK	846.5	2+ 1+	424.25 847.50	3.22E+07 8.46E+05 3.31E+07	16.28
5	AEFVEVTK	921.48	2+ 1+	461.74 922.48	2.57E+07 2.45E+06 2.81E+07	27.42
6	YLYEIAR	926.49	2+ 1+	464.25 927.49	2.98E+07 3.18E+06 3.30E+07	34.41
7	EKVLTSSAR	989.55	3+ 2+ 1+	330.85 495.78 990.55	2.85E+06 1.56E+07 1.06E+05 1.86E+07	11
8	ALKAWSVAR	1000.58	3+ 2+	334.53 501.29	3.49E+06 1.28E+07 1.63E+07	28.55
9	QTALVELLK	1013.61	2+ 1+	507.81 1014.61	4.27E+07 1.13E+06 4.38E+07	42.19
10	YLYEIARR	1082.59	3+ 2+	361.86 542.30	3.38E+06 9.67E+06 1.30E+07	29.9
11	CCTESLVNR	1137.49	2+	569.75	1.03E+07 1.03E+07	21.6
12	KQTALVELLK	1141.71	3+ 2+ 1+	381.57 571.86 1142.71	5.19E+06 3.48E+07 2.67E+05 4.02E+07	37.06
13	LVNELTEFAK	1162.62	2+ 1+	582.31 1163.62	3.53E+07 7.92E+05 3.61E+07	40.16
14	FKDLGEEHFK	1248.61	3+ 2+ 1+	417.20 625.31 1249.61	4.42E+07 2.74E+07 1.22E+05 7.17E+07	26.35
15	HLVDEPQNLIK	1304.71	3+ 2+ 1+	435.90 653.36 1305.71	4.85E+06 6.80E+07 5.15E+05 7.34E+07	32.23
16	SLHTLFGDELCK	1418.69	3+ 2+ 1+	473.90 710.35 1419.69	2.65E+07 5.18E+07 3.50E+05 7.86E+07	39.75
17	RHPEYAVSVLLR	1438.8	3+ 2+	480.60 720.40	5.55E+07 3.34E+07 8.89E+07	35.2
18	YICDNQDTISSK	1442.63	2+ 1+	722.32 1443.63	6.79E+07 5.09E+05 6.84E+07	23.06

no.	Peptide	Molecular weight	Charge	m/z	Peak area	Retention time
19	TCVADESHAGCEK	1462.58	3+ 2+ 1+	488.53 732.29 1463.58	1.30E+07 2.81E+07 2.41E+05 4.13E+07	10.34
20	LGEYGFQNALIVR	1478.79	3+ 2+	493.93 740.40	4.54E+05 1.13E+07 1.18E+07	43.76
21	LKECCDKPLLEK	1531.77	3+ 2+	511.59 766.89	7.06E+07 2.46E+07 9.52E+07	21.27
22	LCVLHEKTPVSEK	1538.81	3+ 2+	513.94 770.41	3.56E+07 1.51E+07 5.07E+07	24.91
23	DAFLGSFLYEYSR	1566.74	2+	784.37	1.37E+06 1.37E+06	51.82
24	KVPQVSTPTLVEVSR	1638.93	3+ 2+	547.31 820.47	1.06E+08 6.60E+07 1.72E+08	34.54
25	RPCFSALTPDETYVPK	1879.91	3+ 2+	627.64 940.96	1.48E+07 7.50E+06 2.23E+07	38.09
26	LKPDPNTLCDEFKADEK	2018.96	3+ 2+	673.99 1010.48	2.42E+07 4.22E+06 2.84E+07	34.8
27	ECCHGDLLECADDRADLAK	2246.94	3+ 2+	749.98 1124.47	1.56E+07 7.11E+05 1.63E+07	32.43

In nanoLC-MS/MS analysis of the BSA dilution series (ranged from 6 to 750 fmol) I plotted peak areas (averaged for five measurements on different days) of all identified peptides against the analyzed amount of the protein (Figure 2.21).

The obtained results indicated that peptide peak areas correlate linear in the given concentration range, and are repeatable (Figure 2.22). The analytical variability of MS signal (RSD) associated with measurements on five consecutive days was typically above 50 % for BSA at 6 fmol, indicating that the acquisition was at the noise level. RSD of MS signal for 12 fmol was about 40% and at higher concentrations in dilution series came below 20% for each peptide. The RSD of retention time for each peptide was less than 3 %.

Table 2.7 represents R^2 values corresponding to linear regression lines obtained for the characterized peptides in the dilution series of the BSA tryptic digest (Table 2.6). The linearity of the peptide peak areas over applied concentration range (of about 2 orders of magnitude) can be expressed in averaged R^2 value of 0.9878.

It should be noted that in some experiments several peptides, especially hydrophobic and those with long peptide sequences showed nonlinear behaviour. They smeared on the applied columns and almost completely disappeared at lower protein concentrations. I tested several columns and concluded that old columns are responsible for this problem.

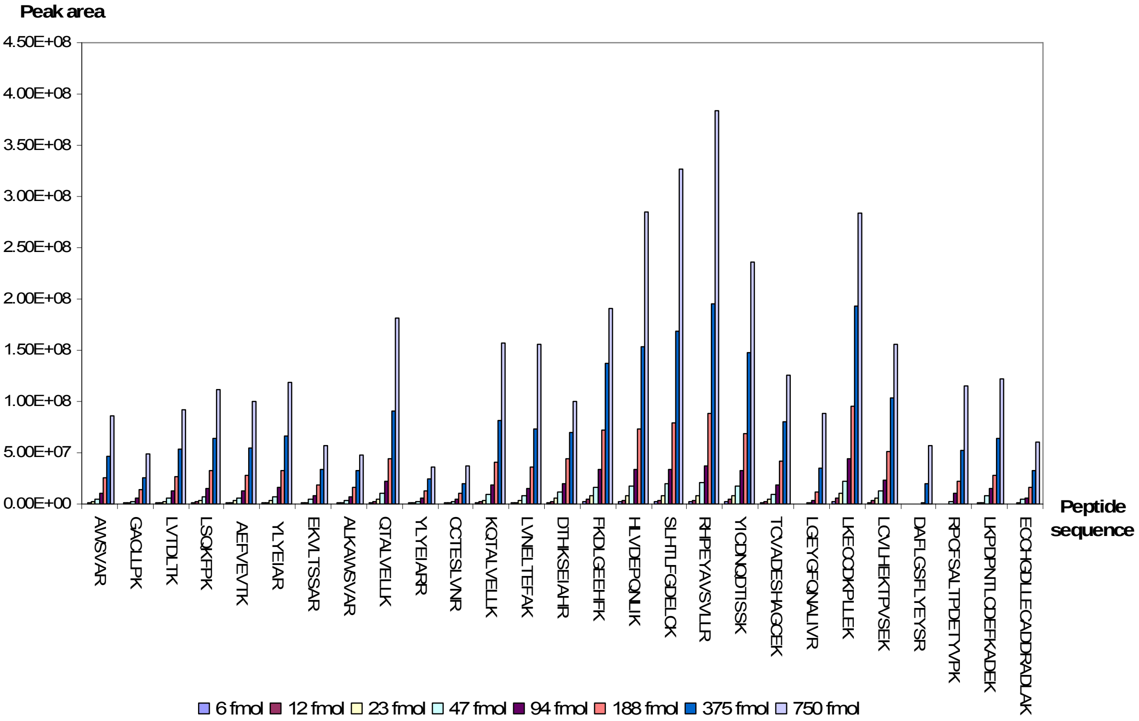
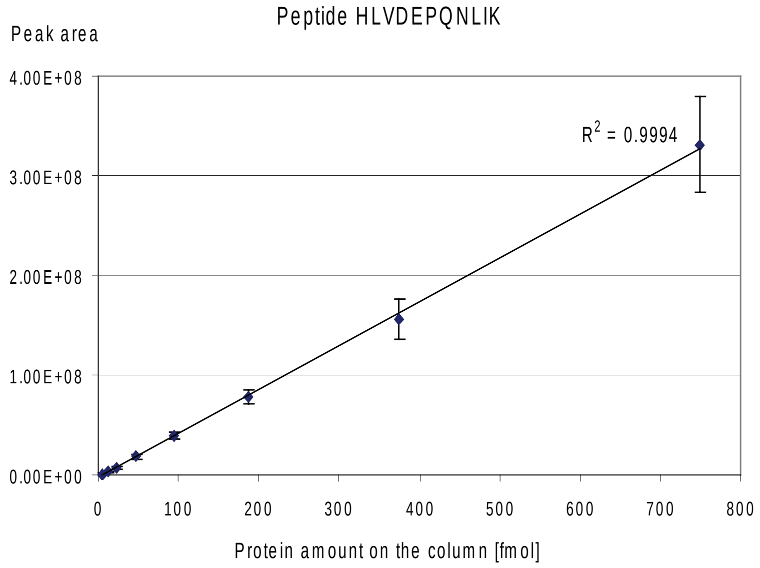


Figure 2.21: Peptides characterized by nanoLC-MS/MS analysis of the dilution series from the BSA tryptic digest.

A bar plot represent the characterized BSA peptides and their corresponding peak areas. The amount of protein loaded on the analytical column ranged from 6 to 750 fmol and is indicated by different colours, as shown in the legend.

It is, therefore, recommendable first to check linear correlation between MS signal of the identified peptides and amount of the analyzed protein for each applied analytical column.

a)



b)

amount of analyzed protein [fmol]	peak area run 1	peak area run 2	peak area run 3	peak area run 4	peak area run 5	average	SD	RSD, %
5.9	5.31E+05	1.40E+06	6.65E+05	6.07E+05	8.47E+05	8.10E+05	3.49E+05	43
11.7	2.17E+06	4.02E+06	2.52E+06	2.53E+06	2.13E+06	2.68E+06	7.72E+05	29
23.4	7.71E+06	9.57E+06	5.49E+06	5.83E+06	7.04E+06	7.13E+06	1.63E+06	23
46.8	1.80E+07	2.01E+07	1.75E+07	1.37E+07	2.01E+07	1.79E+07	2.62E+06	15
93.8	3.39E+07	4.30E+07	3.76E+07	3.87E+07	4.10E+07	3.88E+07	3.46E+06	9
187.5	7.34E+07	8.48E+07	7.11E+07	7.64E+07	8.57E+07	7.83E+07	6.67E+06	9
375	1.54E+08	1.66E+08	1.64E+08	1.21E+08	1.74E+08	1.56E+08	2.10E+07	13
750	2.84E+08	3.56E+08	3.01E+08	3.11E+08	4.02E+08	3.31E+08	4.80E+07	14

Figure 2.22: Correlation between chromatographic peak area and amount of the analyzed proten obtained for BSA peptide HLVDEPQNLIK in the dilution series of the BSA tryptic digest.

The analyzed amount of BSA ranged from 6 to 750 fmol. (a) Averaged peak area (of five measurements) for BSA peptide HLVDEPQNLIK plotted against the corresponding protein amount loaded on the analytical column. A linear curve fit was calculated for the entire dataset (y = 441420x - 3E+06, R^2 = 0.9994) (b) calculated peak areas for each analyzed protein amount obtained in 5 consecutive measurements. Standard and relative standard deviations were calculated for each data point.

Table 2.7: R^2 values corresponding to linear regression lines obtained for the characterized peptides in the BSA dilution series.

no.	Peptide	R^2	no.	Peptide	R^2
1	AWSVAR	0.9959	15	HLVDEPQNLIK	0.9979
2	GACLLPK	0.9985	16	SLHTLFGDELCK	0.9992
3	LVTDLTK	0.9915	17	RHPEYAVSVLLR	0.9991
4	LSQKFPK	0.9941	18	YICDNQDTISSK	0.984
5	AEFVEVTK	0.9974	19	TCVADESHAGCEK	0.9798
6	YLYEIAR	0.9959	20	LGEYGFQNALIVR	0.9874
7	EKVLTSSAR	0.987	21	LKECCDKPLLEK	0.9676
8	ALKAWSVAR	0.9675	22	LCVLHEKTPVSEK	0.9733
9	QTALVELLK	1	23	DAFLGSFLYEYSR	0.9577
10	YLYEIARR	0.9581	24	KVPQVSTPTLVEVSR	0.9977
11	CCTESLVNR	0.998	25	RPCFSALTPDETYVPK	0.9987
12	KQTALVELLK	0.9994	26	LKPDPNTLCDEFKADEK	0.9985
13	LVNELTEFAK	0.9992	27	ECCHGDLLECADDRADLAK	0.9966
14	FKDLGEEHFK	0.95		**Average**	0.9878

b) Study of a five protein mixture

To further evaluate the quantification method for protein profiling of protein digest mixtures I analysed tryptic digest of a five protein-mixture containing myosin (223724 kDa), β-galactosidase (116409 kDa), BSA (69193 kDa), alcohol dehydrogenase (37282 kDa), and myoglobin (16940 kDa). Tryptic digest of the biggest protein myosin resulted in high number of peptides and significantly increased the complexity of the analyzed samples. The amounts of proteins contained in serial dilutions and analyzed by nanoLC-MS/MS are represented in the Table 2.8.

Table 2.8: Proteins contained in the five-protein digest mixture and their amounts analyzed by nanoLC-MS/MS.

Protein	Molecular weight [kDa]	Amount of protein loaded onto the analytical column, [fmol]					
		mixture 1	mixture 2	mixture 3	mixture 4	mixture 5	mixture 6
Myosin	223	18	36	73	145	290	580
b-Galactosidase	116	27	53	106	213	425	850
BSA	69	23	47	94	188	375	750
Alc. Dehydrogenase	37	24	48	96	193	385	770
Myoglobin	17	29	58	116	233	465	930

Figur 2.23 represents a base peak ion chromatogram of the analyzed five-protein digest mixture as well as base peak ion chromatograms of each separately analyzed protein. Altogether more than 200 peptides could be identified by nanoLC-MS/MS analysis here.

From the full scan mass spectra of the analyzed dilution series of the five-protein digest mixture I generated XICs for all BSA peptides characterized before (Table 2.6 and 2.7). Here I aimed to prove whether MS peptide signal in the protein mixture linearly correlates with the analyzed amount of protein. Figure 2.24 shows XICs of differently charged ions of BSA peptide KVPQVSTPTLVEVSR.

The obtained data showed that the peak areas of the peptides obtained from the dilution series of the five-protein digest mixture linearly correlated to the amounts of the analyzed proteins. Linear regression R^2 values for the BSA peptides characterized from the five-protein digest mixture (as previously characterized in the BSA digest, Table 2.7) are shown in the Table 2.9. The linearity within dilution range was here in average 0.9908. The variability of the MS signal for a single peptide (measured in three runs) slightly increased in the protein mixture compared to the analysis of the single protein but not exceeded 30% in the tested concentration range (from about 20 to 1000 fmol).

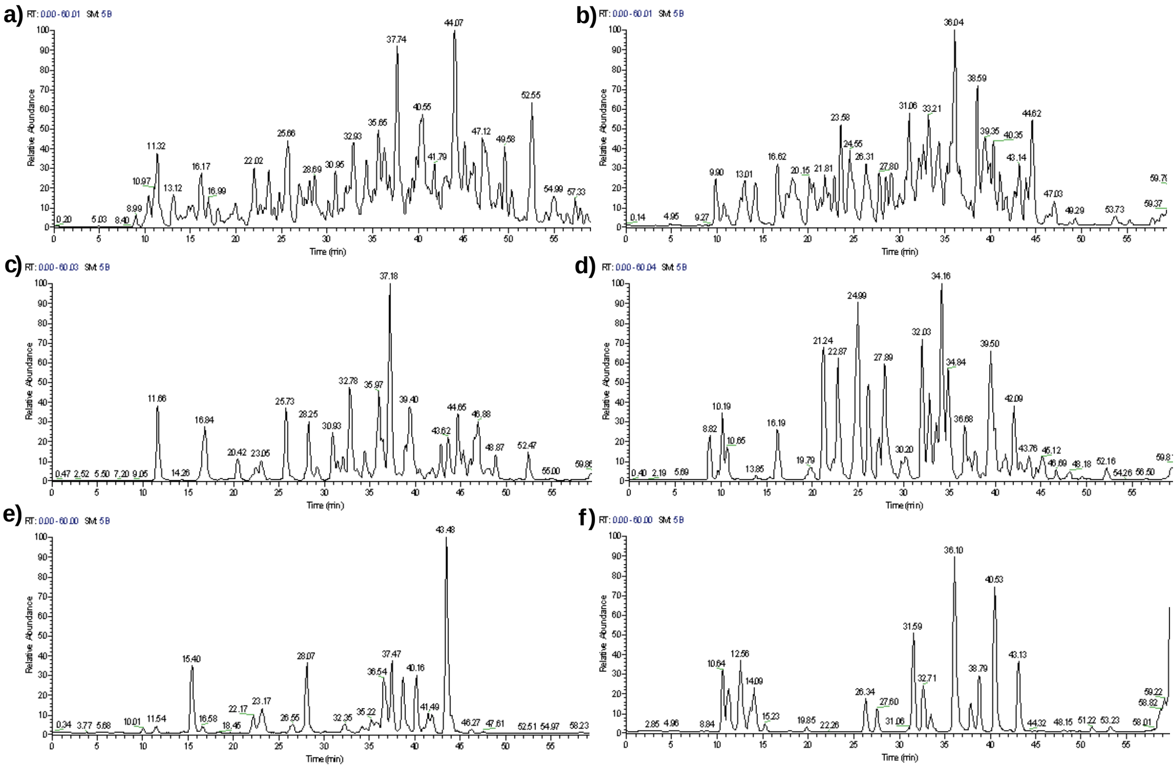


Figure 2.23: Base peak ion chromatogram of the analyzed five-protein digest mixture as well as base peak ion chromatograms of each separately analyzed protein.

a) Base peak ion chromatogram of the five-protein digest mixture; (b), (c), d), (e), (f) base peak ion chromatograms of the tryptic digests of myosin (290 fmol), β-galactosidase (425 fmol), BSA (345 fmol), alcohol dehydrogenase (385 fmol) and myoglobin (465 fmol), respectively.

Table 2.9: R^2 values corresponding to linear regression lines obtained for the characterized BSA peptides from the five-protein digest mixture.

no.	Peptide	R^2	no.	Peptide	R^2
1	AWSVAR	0.9989	15	HLVDEPQNLIK	0.9888
2	GACLLPK	0.9988	16	SLHTLFGDELCK	0.9966
3	LVTDLTK	0.993	17	RHPEYAVSVLLR	0.9887
4	LSQKFPK	0.999	18	YICDNQDTISSK	0.9866
5	AEFVEVTK	0.9974	19	TCVADESHAGCEK	0.9872
6	YLYEIAR	0.9992	20	LGEYGFQNALIVR	0.991
7	EKVLTSSAR	0.9756	21	LKECCDKPLLEK	0.9808
8	ALKAWSVAR	0.9725	22	LCVLHEKTPVSEK	0.9979
9	QTALVELLK	0.9922	23	DAFLGSFLYEYSR	0.9935
10	YLYEIARR	0.9798	24	KVPQVSTPTLVEVSR	0.9962
11	CCTESLVNR	0.993	25	RPCFSALTPDETYVPK	0.9975
12	KQTALVELLK	0.9995	26	LKPDPNTLCDEFKADEK	0.99
13	LVNELTEFAK	0.9968	27	ECCHGDLLECADDRADLAK	0.9732
14	FKDLGEEHFK	0.9878		**Average**	0.9908

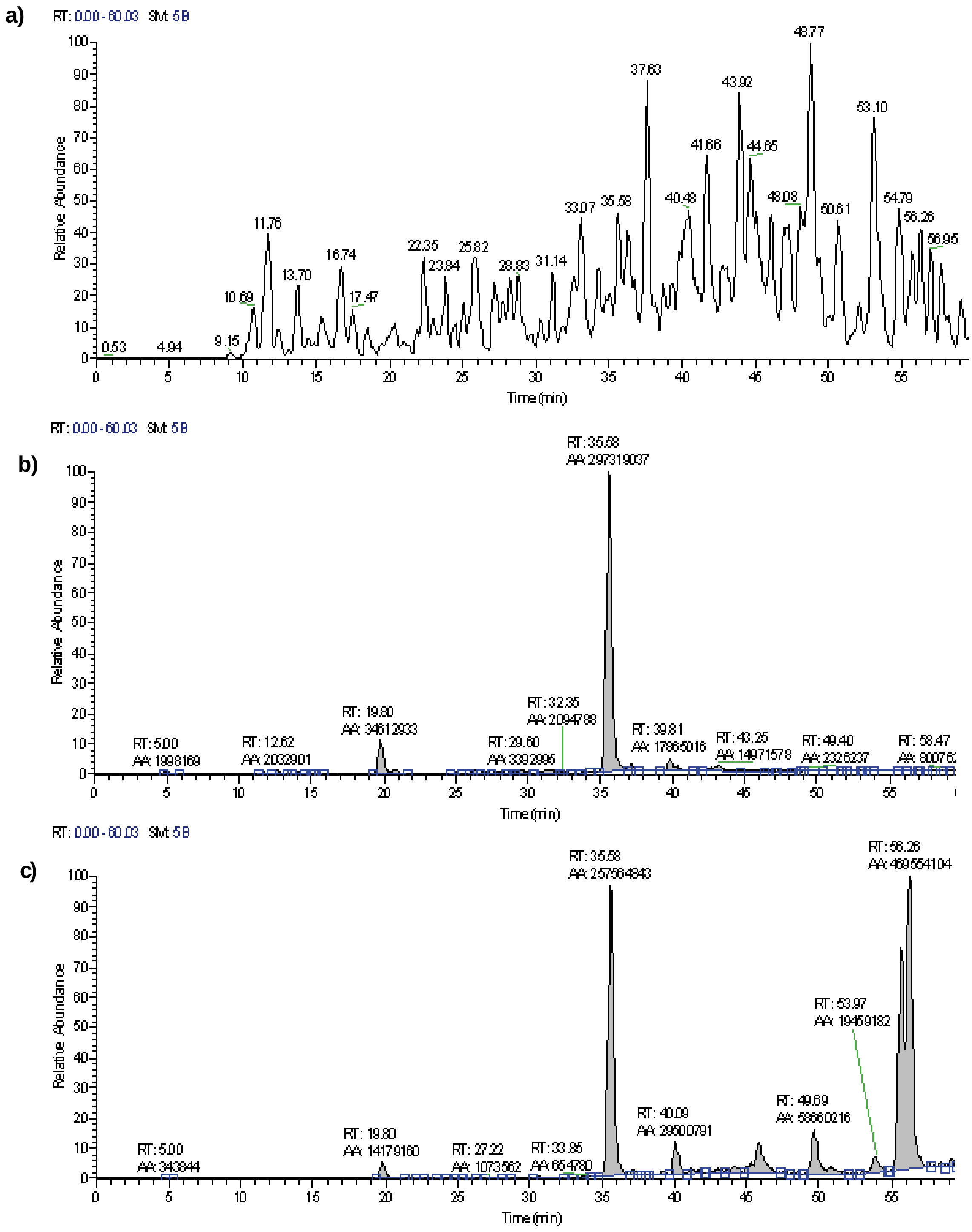


Figure 2.24: Base peak ion chromatogram of the five-protein digest mixture and extracted ion chromatograms (XICs) of the differently charhed ions of BSA peptide KVPQVSTPTLVEVSR.

a) Base peak ion chromatogram of the five-protein digest mixture, containing myosin (290 fmol), β-galactosidase (425 fmol), BSA (345 fmol), alcohol dehydrogenase (385 fmol) and myoglobin (465 fmol). (b), (c), (d) XICs of triple, double and single charged ions of the BSA peptide KVPQVSTPTLVEVSR, respectively. Scan numbers corresponding to the MS/MS spectra of this peptide helped to identify its correct retention time, at 35.58 min.

Proteins present in the five-protein digest mixture were analyzed separately; their amounts contained in serial dilutions were the same as depicted in the Table 2.8. I plotted linear curves for some peptides analyzed from the single protein digest (β-galactosidase, BSA, alcohol dehydrogenase and myoglobin) and from the five-protein digest mixtures (Figure 2.25). The results showed consistency between peptide peak intensities as well as linearity in the analysis of the single proteins and of the protein mixture.

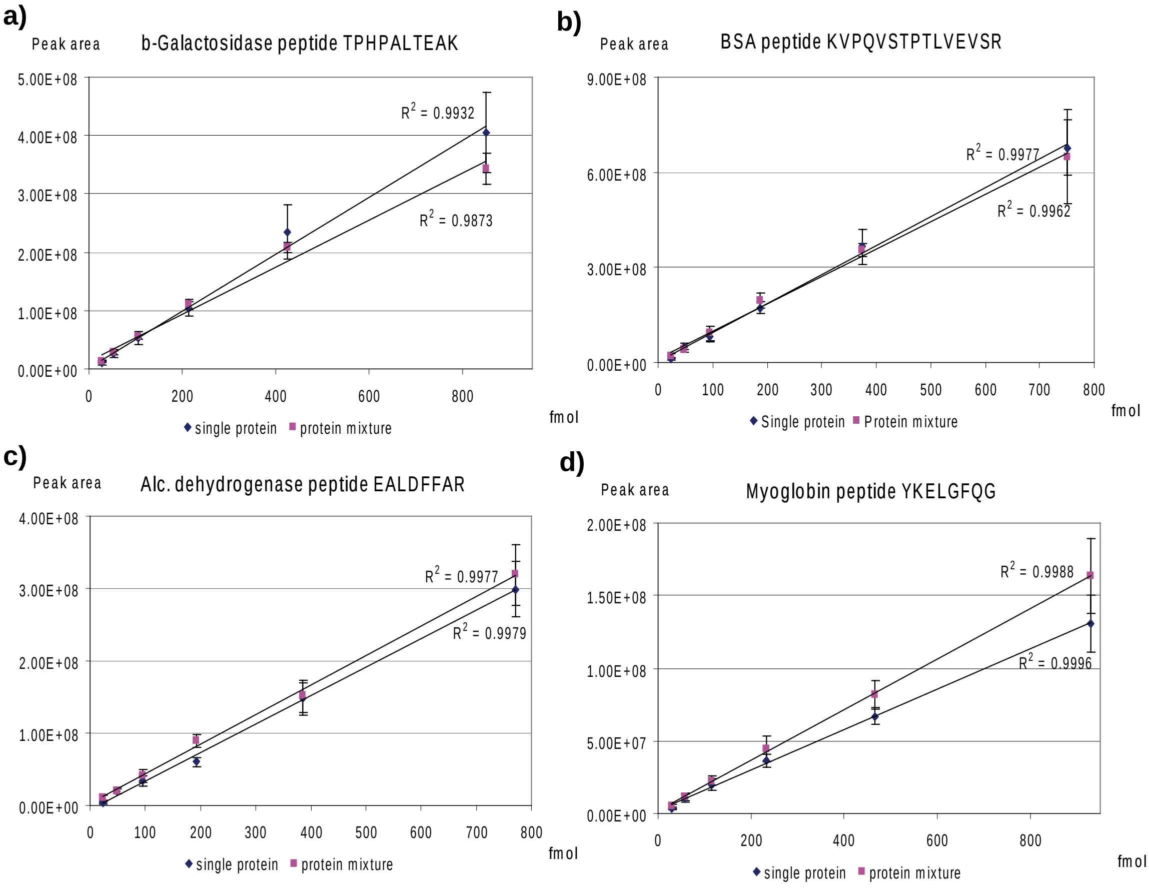


Figure 2.25: Linear curves plotted for β-galactosidase, BSA, alcohol dehydrogenase, and myoglobin peptides analyzed from the single protein digest and from the five-protein digest mixtures.

The amounts of proteins contained in serial dilutions (of the single protein digest and of the five-protein digest mixtures) and analyzed by nanoLC-MS/MS are represented in the Table 2.8. Linear regression lines were obtained for β-Galactosidase peptide TPHPALTEAK (a), BSA peptide KVPQVSTPTLVEVSR (b), alcohol dehydrogenase peptide EALDFFAR (c) and myoglobin peptide YKELGFQG (d).

2.1.4.2 Application of this approach for absolute quantification of proteins

The study presented above demonstrated that peptide peak areas from nanoLC-MS/MS analysis can be used for quantitative protein analysis in relatively simple mixtures such as gel bands (spots) from one- or two-dimensional electrophoresis. This method appeared to be accurate and reproducible.

The applicability of this quantitation approach to measure changes in relative protein concentration even in complex samples such as digest of total human plasma protein has been proved by several research groups [272-276].

I presumed that this method might also be useful for absolute quantification of simple protein digest mixtures, when applying calibrating curves of the reference peptides obtained by in-solution digestion of known amounts of the corresponding protein standards (assuming that the recovery of in-solution digestion is close to 100%).

I set out to employ this method in quantification of in-gel digestion products in the kinetic study previously described in chapter 2.1.3.4. This would provide independent information to the quantitative study based on ^{18}O-labeled peptides and MALDI MS.

2.1.4.3 Kinetic study of accelerated in-gel digestion of proteins by glycosylated trypsins

As in previously described kinetic experiments I monitored the time course of the peptide yield observed by in-gel digestion of a standard protein (BSA) by glycosylated trypsins at accelerated conditions.

Gel bands containing 1 pmol BSA were digested for 0.5, 1.5 and 3 h at 55°C by MAT-BT and RAF-BT applied at the concentration of 1.4 and 2.8 µM. The obtained peptides were extracted from the gel matrix and dried down. The peptide mixture was redissolved in 10 µL of 0.05% TFA and 2 µL of the sample were analyzed by nanoLC-MS/MS. For each peptide subjected for quantification a calibration curve of the corresponding peptide was generated, which was obtained from serial dilutions of in-solution digest of BSA (chapter 4.1.3.7).

It should be noted that analysis even of the same sample results in differences in the peak areas of the peptides from one run to other. This may be caused by experiment dependent parameters such as differences in sample preparation (pipetting errors, incomplete digestion) or instrument dependent parameters such as errors in sample injection, HPLC or MS instrument performance. To get better statistics of the experiments each sample as well as each analysis was prepared in triplicate.

The calculated digestion yields were compared with those of conventional digestion (37 °C, overnight, by native BT at concentration ~ 0.5 μM) and accelerated digestion developed by Havliš et al. [2] (55°C, digestion time ranging from 0.5 to 1 h, by methylated porcine trypsin at the concentration ~ 1.5 μM) (Figure 2.26). In contrast to the quantification method based on stable isotope labeling of peptides with ^{18}O and MALDI TOF MS label-free approach based on nano-LC-MS/MS analysis allowed me to use more reference peptides for quantification experiments. In addition higher dynamic range of detection and ability to analyse complex protein mixtures employed in this technology enabled application of enzymes at high concentration. To evaluate the digestion yield amounts of the following six peptides were averaged: AEFVEVTK (M = 921.48), YLYEIAR (M = 926.49), KQTALVELLK (M = 1141.71), LVNELTEFAK (M = 1162.62), HLVDEPQNLIK (M = 1304.71), KVPQVSTP-TLVEVSR (M = 1638.93).

The peptide recovery of BSA obtained upon 30 min of in-gel digestion by MAT-BT and RAF-BT at the concentration of 1.4 μM gained 38 and 39 % of conventional digestion yield, respectively (Table 2.10). This recovery was higher than those determined in former kinetic study based on ^{18}O-labeled peptides and MALDI TOF analysis (13 and 19 % of CD yield for MAT-BT and RAF-BT applied at the enzyme concentration of 0.98 and 0.86 μM, respectively) (Table 2.3). These results can be explained by quantification errors present in both experiments. I also reasoned that the yield of in-gel digestion was underestimated by ^{18}O-labeling quantification method, since less reference peptides were applied here compared to the label-free approach.

On the other hand, 30 min of in-gel digestion of BSA using both conjugates at the enzyme concentration of 2.8 μM resulted in a peptide recovery of 43 and 44 % (determined by label-free approach), well in agreement with results obtained in previous quantification experiments (44 % recovery for MAT-BT applied at the concentration of 2.45 μM, and 39 % recovery for RAF-BT applied at the concentration of 2.2 μM).

Between 68 and 71% of conventional digestion yield was reached upon 3 hours of in-gel digestion of BSA by trypsin conjugates at the tested concentrations (Table 2.11). These results were consistent with results obtained in previous quantification experiments (54 and 62% recovery for MAT-BT applied at the concentration of 0.98

and 2.45 µM, respectively, and 49 and 58% recovery for RAF-BT applied at the concentration of 0.86 and 2.2 µM, respectively).

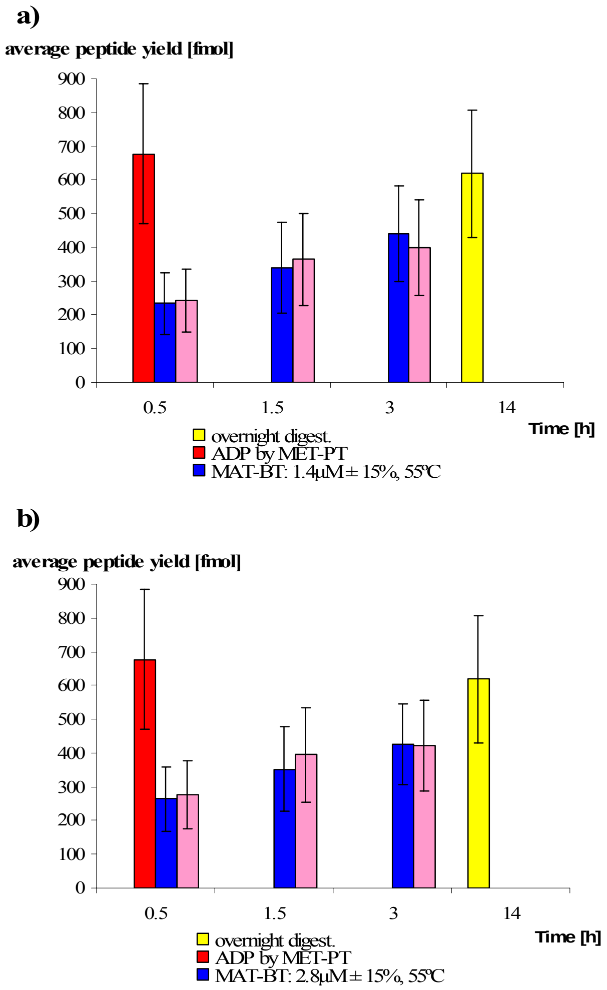


Figure 2.26: Time course of the averaged peptide yield observed by in-gel digestion of BSA using MAT-BT and RAF-BT at accelerated conditions.

Applied enzyme concentrations: a) 1.4 µM and b) 2.8 µM.

BSA bands (1 pmol) were in-gel digested by MAT-BT and RAF-BT at 55°C and an enzyme concentration 1.4 (a) and 2.8 µM (b); digestion times were: 30, 1.5h and 3h. The blue and pink coloured bars represent the recovery of accelerated digestion by MAT-BT and RAF-BT, respectively. The red and yellow coloured bars represent the recovery of accelerated digestion by MET-PT (55°C, 0.5 to 1 h, MET-PT at concentration ~ 1.5 µM) and conventional digestion (37°C, overnight, BT at concentration ~ 0.5 µM), respectively.

Interestingly, the digestion yield could not be improved by increasing concentration of trypsin conjugates (in contrast to the former kinetic study). Taking into account that a typical band of 12% polyacrylamide gel (with approximate size 0.8 mm x 0.8 mm x 6.4 mm) absorbs 4 µL of digestion buffer [2], a gel band contacting 1 pmol of

BSA would result in the initial protein concentration of 0.26 µM, if reaction occurs in-solution. According to the reported Km of 1.6 ± 0.2 µM µM for trypsin-catalyzed protein cleavage in-solution [301], at this protein concentration trypsin is not saturated with substrate. These results rather confirm the assumption of poor accessibility of the in gel matrix imbedded substrate for the bulky molecules of glycosylated trypsins.

Table 2.10: : Averaged peptide yield of BSA obtained upon 30 min of accelerated in-gel digestion by MAT-BT and RAF-BT at enzyme concentration 1.4 and 2.8 µM.

Modified enzyme	Concentration [µM] [a]	Yield [b] [fmol]	SD [c]	RSD [d]	Recovery [%] [e]
native BT [f]	0.5	618	189	31	
MET-PT [g]	1.5	678	209	31	110
MAT-BT	1.4	234	93	40	38
	2.8	263	94	36	43
RAF-BT	1.4	244	94	39	39
	2.8	276	100	36	45

a) Concentration determined by amino acid analysis; b) Average peptide yield of the digestion; c) Standard deviation of the calculated peptide yield; d) Relative standard deviation of the calculated peptide yield; e) Percentage of the tryptic peptide recovery of conventional digestion; f) Conventional digestion (overnight, at 37 °C); g) accelerated digestion protocol [2].

Table 2.11: Averaged peptide yield of BSA obtained upon 3 hours of accelerated in-gel digestion by MAT-BT and RAF-BT at enzyme concentration 1.4 and 2.8 µM.

Modified enzyme	Concentration [µM] [a]	Yield [b] [fmol]	SD [c]	RSD [d]	Recovery [%] [e]
native BT [f]	0.5	618	189	31	
MAT-BT	1.4	441	140	32	71
	2.8	426	120	28	69
RAF-BT	1.4	400	141	35	65
	2.8	422	136	32	68

a) Concentration determined by amino acid analysis; b) Average peptide yield of the digestion; c) Standard deviation of the calculated peptide yield; d) Relative standard deviation of the calculated peptide yield; e) Percentage of the tryptic peptide recovery of conventional digestion; f) Conventional digestion (overnight, at 37 °C).

Further I set out to investigate whether digestion of proteins at higher temperature would accelerate the protein cleavage. The tested digestion temperature was set at 65°C and in-gel digestion was performed for 30 min at the enzyme concentration of trypsin conjugates of 2.8 µM. Figure 2.27 represents the peptide recovery obtained by in-gel digestion of BSA by MAT-BT (Figure 2.27 a) and RAF-BT (Figure 2.27 b) at 55 and 65°C. No significant changes were observed at the tested digestion conditions. Digestion recovery could not be improved by increasing incubation temperature.

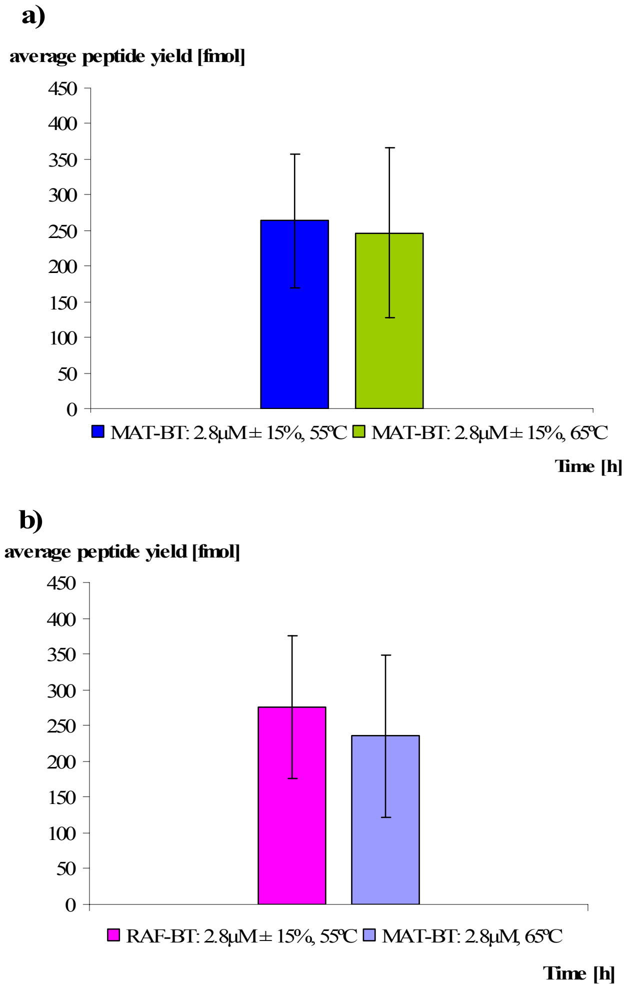
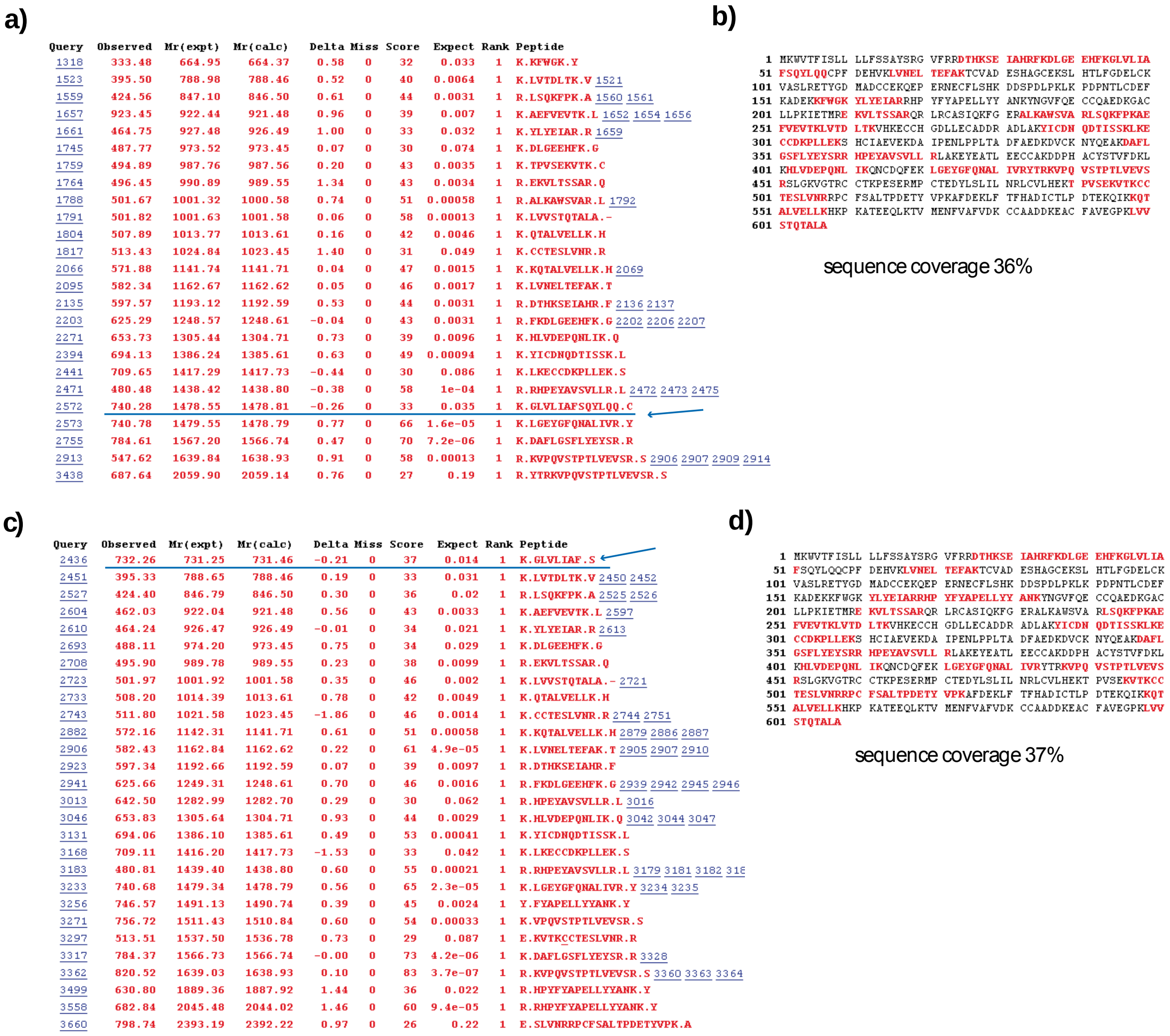


Figure 2.27: Averaged peptide recovery obtained by in-gel digestion of BSA by MAT-BT and RAF-BT at different temperatures.

BSA bands (1 pmol) were in-gel digested by MAT-BT (a) and RAF-BT (b) at 55 and 65°C and at an enzyme concentration of 2.8 µM; the digestion was performed for 30 min. The blue and pink coloured bars represent the recovery of digestion obtained by MAT-BT and RAF-BT at 55°C, respectively. The green and pale blue coloured bars represent the recovery of digestion obtained by MAT-BT and RAF-BT at 65°C, respectively.

MS/MS information available by the applied technique (in contrast to peptides mass fingerprinting by MALDI TOF MS) enabled me to compare the cleavage specifity of trypsin conjugates and native trypsin. To this end I performed database searches without restricting the enzyme cleavage specifity in order to see whether some non-tryptic peptides were produced during the digestion. Figure 2.28 represents peptides identified upon MASCOT searches from nano-LC-MS/MS data of BSA in-gel digests obtained by accelerated digestion using MAT-BT (55°C, 3h of digestion) and by conventional digestion using native bovine trypsin (37°C, overnight digestion).

With exception of two half-tryptic peptides GLVIAFSQYLQQ (obtained by MASCOT searches of LC-MS/MS data from an accelerated in-gel digest of BSA by MAT-BT) and GLVIAFS (obtained by MASCOT searches of LC-MS/MS data from a conventional in-gel digest of BSA by native BT) all fragmented peptides were fully tryptic, confirming unchanged cleavage specifity of trypsin conjugates. Both half-tryptic peptides are probably derived from orifice fragmentation of the long tryptic peptide GLVIAFSQYLQQCPFDEHVK. The number of identified peptides in accelerated in-gel digestion was 22, (covering 36 % of BSA sequence) in conventional digestion 28 (covering 37% of BSA sequence).



a)

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
1318	333.48	664.95	664.37	0.58	0	32	0.033	1	K.KFWGK.Y
1523	395.50	788.98	788.46	0.52	0	40	0.0064	1	K.LVTDLTK.V 1521
1559	424.56	847.10	846.50	0.61	0	44	0.0031	1	R.LSQKFPK.A 1560 1561
1657	923.45	922.44	921.48	0.96	0	39	0.007	1	K.AEFVEVTK.L 1652 1654 1656
1661	464.75	927.48	926.49	1.00	0	33	0.032	1	K.YLYEIAR.R 1659
1745	487.77	973.52	973.45	0.07	0	30	0.074	1	K.DLGEEHFK.G
1759	494.89	987.76	987.56	0.20	0	43	0.0035	1	K.TPVSEKVTK.C
1764	496.45	990.89	989.55	1.34	0	43	0.0034	1	R.EKVLTSSAR.Q
1788	501.67	1001.32	1000.58	0.74	0	51	0.00058	1	R.ALKAWSVAR.L 1792
1791	501.82	1001.63	1001.58	0.06	0	58	0.00013	1	K.LVVSTQTALA.-
1804	507.89	1013.77	1013.61	0.16	0	42	0.0046	1	K.QTALVELLK.H
1817	513.43	1024.84	1023.45	1.40	0	31	0.049	1	K.CCTESLVNR.R
2066	571.88	1141.74	1141.71	0.04	0	47	0.0015	1	K.KQTALVELLK.H 2069
2095	582.34	1162.67	1162.62	0.05	0	46	0.0017	1	K.LVNELTEFAK.T
2135	597.57	1193.12	1192.59	0.53	0	44	0.0031	1	R.DTHKSEIAHR.F 2136 2137
2203	625.29	1248.57	1248.61	-0.04	0	43	0.0031	1	R.FKDLGEEHFK.G 2202 2206 2207
2271	653.73	1305.44	1304.71	0.73	0	39	0.0096	1	K.HLVDEPQNLIK.Q
2394	694.13	1386.24	1385.61	0.63	0	49	0.00094	1	K.YICDNQDTISSK.L
2441	709.65	1417.29	1417.73	-0.44	0	30	0.086	1	K.LKECCDKPLLEK.S
2471	480.48	1438.42	1438.80	-0.38	0	58	1e-04	1	R.RHPEYAVSVLLR.L 2472 2473 2475
2572	740.28	1478.55	1478.81	-0.26	0	33	0.035	1	K.GLVLIAFSQYLQQ.C
2573	740.78	1479.55	1478.79	0.77	0	66	1.6e-05	1	K.LGEYGFQNALIVR.Y
2755	784.61	1567.20	1566.74	0.47	0	70	7.2e-06	1	K.DAFLGSFLYEYSR.R
2913	547.62	1639.84	1638.93	0.91	0	58	0.00013	1	R.KVPQVSTPTLVEVSR.S 2906 2907 2909 2914
3438	687.64	2059.90	2059.14	0.76	0	27	0.19	1	R.YTRKVPQVSTPTLVEVSR.S

b)

1 MKWVTFISLL LLFSSAYSRG VFRR**DTHKSE IAHRFKDLGE EHFKGLVLIA**

51 **FSQYLQQ**CPF DEHVK**LVNEL TEFAK**TCVAD ESHAGCEKSL HTLFGDELCK

101 VASLRETYGD MADCCEKQEP ERNECFLSHK DDSPDLPKLK PDPNTLCDEF

151 KADEK**KFWGK YLYEIAR**RHP YFYAPELLYY ANKYNGVFQE CCQAEDKGAC

201 LLPKIETMR**E KVLTSSAR**QR LRCASIQKFG ER**ALKAWSVA RLSQKFPKAE**

251 **FVEVTKLVTD LTK**VHKECCH GDLLECADDR ADLAK**YICDN QDTISSKLKE**

301 **CCDKPLLEK**S HCIAEVEKDA IPENLPPLTA DFAEDKDVCK NYQEAK**DAFL**

351 **GSFLYEYSRR HPEYAVSVLL R**LAKEYEATL EECCAKDDPH ACYSTVFDKL

401 K**HLVDEPQNL IK**QNCDQFEK **LGEYGFQNAL IVRYTRKVPQ VSTPTLVEVS**

451 **R**SLGKVGTRC CTKPESERMP CTEDYLSLIL NRLCVLHEK**T PVSEKVTKCC**

501 **TESLVNR**RPC FSALTPDETY VPKAFDEKLF TFHADICTLP DTEKQIK**KQT**

551 **ALVELLK**HKP KATEEQLKTV MENFVAFVDK CCAADDKEAC FAVEGPK**LVV**

601 **STQTALA**

sequence coverage 36%

c)

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
2436	732.26	731.25	731.46	-0.21	0	37	0.014	1	K.GLVLIAF.S
2451	395.33	788.65	788.46	0.19	0	33	0.031	1	K.LVTDLTK.V 2450 2452
2527	424.40	846.79	846.50	0.30	0	36	0.02	1	R.LSQKFPK.A 2525 2526
2604	462.03	922.04	921.48	0.56	0	43	0.0033	1	K.AEFVEVTK.L 2597
2610	464.24	926.47	926.49	-0.01	0	34	0.021	1	K.YLYEIAR.R 2613
2693	488.11	974.20	973.45	0.75	0	34	0.029	1	K.DLGEEHFK.G
2708	495.90	989.78	989.55	0.23	0	38	0.0099	1	R.EKVLTSSAR.Q
2723	501.97	1001.92	1001.58	0.35	0	46	0.002	1	K.LVVSTQTALA.- 2721
2733	508.20	1014.39	1013.61	0.78	0	42	0.0049	1	K.QTALVELLK.H
2743	511.80	1021.58	1023.45	-1.86	0	46	0.0014	1	K.CCTESLVNR.R 2744 2751
2882	572.16	1142.31	1141.71	0.61	0	51	0.00058	1	K.KQTALVELLK.H 2879 2886 2887
2906	582.43	1162.84	1162.62	0.22	0	61	4.9e-05	1	K.LVNELTEFAK.T 2905 2907 2910
2923	597.34	1192.66	1192.59	0.07	0	39	0.0097	1	R.DTHKSEIAHR.F
2941	625.66	1249.31	1248.61	0.70	0	46	0.0016	1	R.FKDLGEEHFK.G 2939 2942 2945 2946
3013	642.50	1282.99	1282.70	0.29	0	30	0.062	1	R.HPEYAVSVLLR.L 3016
3046	653.83	1305.64	1304.71	0.93	0	44	0.0029	1	K.HLVDEPQNLIK.Q 3042 3044 3047
3131	694.06	1386.10	1385.61	0.49	0	53	0.00041	1	K.YICDNQDTISSK.L
3168	709.11	1416.20	1417.73	-1.53	0	33	0.042	1	K.LKECCDKPLLEK.S
3183	480.81	1439.40	1438.80	0.60	0	55	0.00021	1	R.RHPEYAVSVLLR.L 3179 3181 3182 318
3233	740.68	1479.34	1478.79	0.56	0	65	2.3e-05	1	K.LGEYGFQNALIVR.Y 3234 3235
3256	746.57	1491.13	1490.74	0.39	0	45	0.0024	1	Y.FYAPELLYYANK.Y
3271	756.72	1511.43	1510.84	0.60	0	54	0.00033	1	K.VPQVSTPTLVEVSR.S
3297	513.51	1537.50	1536.78	0.73	0	29	0.087	1	E.KVTKCCTESLVNR.R
3317	784.37	1566.73	1566.74	-0.00	0	73	4.2e-06	1	K.DAFLGSFLYEYSR.R 3328
3362	820.52	1639.03	1638.93	0.10	0	83	3.7e-07	1	R.KVPQVSTPTLVEVSR.S 3360 3363 3364
3499	630.80	1889.36	1887.92	1.44	0	36	0.022	1	R.HPYFYAPELLYYANK.Y
3558	682.84	2045.48	2044.02	1.46	0	60	9.4e-05	1	R.RHPYFYAPELLYYANK.Y
3660	798.74	2393.19	2392.22	0.97	0	26	0.22	1	E.SLVNRRPCFSALTPDETYVPK.A

d)

1 MKWVTFISLL LLFSSAYSRG VFRR**DTHKSE IAHRFKDLGE EHFKGLVLIA**

51 **F**SQYLQQCPF DEHVK**LVNEL TEFAK**TCVAD ESHAGCEKSL HTLFGDELCK

101 VASLRETYGD MADCCEKQEP ERNECFLSHK DDSPDLPKLK PDPNTLCDEF

151 KADEKKFWGK **YLYEIARRHP YFYAPELLYY ANK**YNGVFQE CCQAEDKGAC

201 LLPKIETMR**E KVLTSSAR**QR LRCASIQKFG ERALKAWSVA R**LSQKFPKAE**

251 **FVEVTKLVTD LTK**VHKECCH GDLLECADDR ADLAK**YICDN QDTISSKLKE**

301 **CCDKPLLEK**S HCIAEVEKDA IPENLPPLTA DFAEDKDVCK NYQEAK**DAFL**

351 **GSFLYEYSRR HPEYAVSVLL R**LAKEYEATL EECCAKDDPH ACYSTVFDKL

401 K**HLVDEPQNL IK**QNCDQFEK **LGEYGFQNAL IVR**YTR**KVPQ VSTPTLVEVS**

451 **R**SLGKVGTRC CTKPESERMP CTEDYLSLIL NRLCVLHEKT PVSE**KVTKCC**

501 **TESLVNRRPC FSALTPDETY VPK**AFDEKLF TFHADICTLP DTEKQIK**KQT**

551 **ALVELLK**HKP KATEEQLKTV MENFVAFVDK CCAADDKEAC FAVEGPK**LVV**

601 **STQTALA**

sequence coverage 37%

Figure 2.28: Peptides identified upon MASCOT database searches from nano-LC-MS/MS data of BSA in-gel digests obtained by accelerated digestion using MAT-BT and by conventional digestion using native bovine trypsin.

(a) and (c) peptides identified upon MASCOT database searches from nano-LC-MS/MS data of BSA in-gel digests obtained by accelerated and conventional digestion, respectively; (b) and (d) sequence coverage of identified peptides within BSA protein sequence for data obtained by accelerated and conventional digestion, respectively.

Gel bands containing 1 pmol BSA were in-gel digested using accelerated conditions by MAT-BT (55°C; 3 h, at an enzyme concentration of 1.4 µM) and conventional digestion (37°C; overnight, at an enzyme concentration of 0.5 µM); Database searching was performed using mass tolerance for precursor and fragment ions of 2.0 and 0.5 Da, respectively; oxidation of methionine was considered as a variable modification.

2.1.4.4 Conclusion on the performed kinetic study

Application of two different quantification approaches in the described kinetic study allowed me to evaluate the catalytic efficiency of the tested trypsin conjugates in accelerated in-gel digestion of proteins. Both quantification studies were in general consistent and showed that the recovery of conventional in-gel digestion could not be improved using glycosylated trypsins. At the best between 60 to 70% of conventional digestion yield could be achieved upon accelerated in-gel digestion of BSA by MAT and RAF-BT. The outcome of the performed experiments prompted me to conclude that sterically hindered enzyme/substrate binding (due to the bulky structure of glycosylated trypsins) is the main factor responsible for the reduced efficiency of trypsin conjugates in in-gel digestion of proteins.

2.2 Validations of protein identifications with borderline statistical confidence

In first part of my work I focused on the improvement of the conventional in-gel digestion protocol in order to simplify sample preparation and increase the efficiency of digestion. This goal, however, could not be achieved by application of glycosylated trypsin conjugates.

Another important issue in bottom-up proteomics is reliability of the protein identifications. This problem derives from the high complexity of the analyzed protein mixtures and limited sensitivity and dynamic range of the common analytical instruments. Thus, the identification of large number of proteins relies on matching one or two spectra of marginal quality, yielding protein identifications with borderline statistical confidence. These borderline protein identifications include false positive and false negative hits. How can we distinguish false hits from true? Probabilistic scoring, which is applied in a variety of search engines such as MASCOT [170], Sequest [172] etc. only suggest the threshold of statistically reliable assignments. To answer this question I set out to develop fast and reliable method for validation of borderline hits, which complements conventional database searching and can be applied in large scale proteomic analysis.

2.2.1 Combination of *de novo* sequencing (PepNovo) and MS BLAST searches for independent validation of database searching hits

In contrast to conventional database searching *de novo* sequencing algorithms read out peptide sequences directly from fragment ion spectra independently of available sequence resources [159, 177-182]. Since *de novo* interpretation of tandem mass spectra results in many sequence proposals, which are highly redundant and error-prone, we proposed to combine it with sequence-similarity searching tool MS BLAST [220, 224], which tolerates redundancy and partial inaccuracy of candidate peptides and employs an independent scoring scheme. The combination of *de novo* sequencing and MS BLAST would provide a cross-validation tool for obtained database searching hits.

De novo sequencing program PepNovo developed by Frank et al. [179] has been reported to have good quality predictions and transparent internal quality score.

Moreover it is fast and can be interfaced to MS BLAST. To assess if a combination of *de novo* program PepNovo and MS BLAST could validate MASCOT hits with marginal ions scores, I composed a dataset comprising 100 high-quality tandem mass spectra that unequivocally matched sequences of full tryptic peptides in a database. In each spectrum the actual signal-to-noise level was in *silico* altered by gradually decreasing the intensities of matching peaks, while the abundance of peaks of chemical noise was fixed (Figure 2.29). So we simulated the situation, where the protein identification only relies on matching a single spectrum of marginal quality. Each series of spectra with perturbed signal-to-noise ratios was subjected, in parallel, to MASCOT searches and *de novo* interpretation by PepNovo software. Up to seven sequence candidates per each interpreted spectrum were merged into a query string, which was then submitted to MS BLAST search. The whole procedure was done using a script written by Henrik Thomas (Shevchenko Group, MPI-CBG).

Within each series, I aimed to determine the MASCOT ions score and PepNovo quality score for the two spectra having the lowest signal-to-noise ratios, whose PepNovo sequencing and MS BLAST searching either confidently identified (according to MS BLAST scoring scheme) the correct peptide in a comprehensive database or listed the correct peptide among the top 50 nonconfident hits in the MS BLAST output (Figure 2.30). However, in several cases, such spectra were not identified (altogether six peptide sequences). On several occasions, PepNovo/MS BLAST failed to match the expected sequence by interpreting even the initial high-quality spectrum, or the expected peptide was missing among nonconfident hits in the MS BLAST output. Therefore, the actual number of data points in Figures built using this dataset (Figures 2.31 and 2.32) was less than the expected 100.

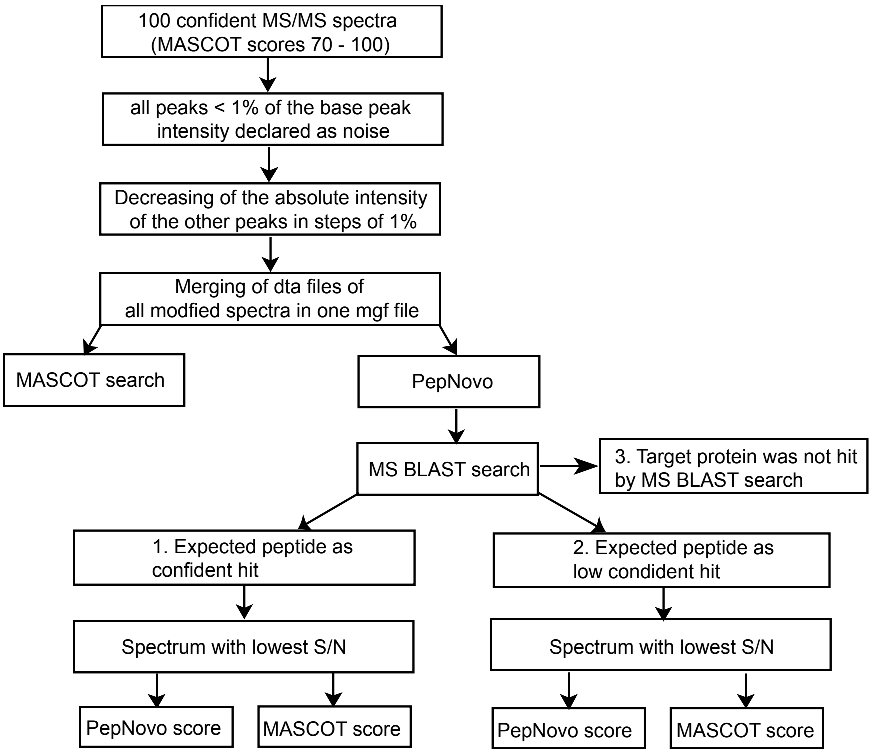
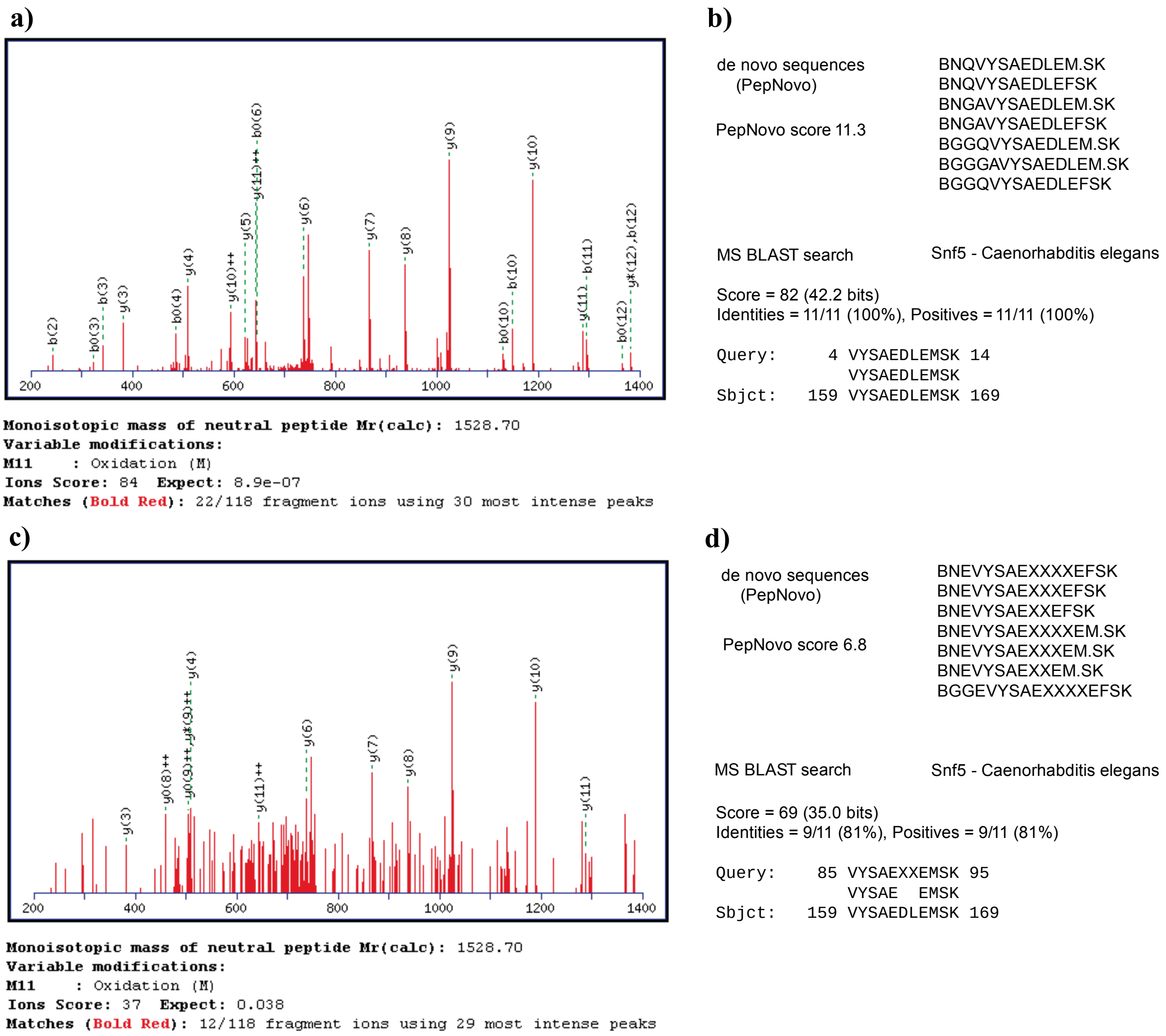

100 confident MS/MS spectra (MASCOT scores 70 - 100)
all peaks < 1% of the base peak intensity declared as noise
Decreasing of the absolute intensity of the other peaks in steps of 1%
Merging of dta files of all modfied spectra in one mgf file
MASCOT search
PepNovo
MS BLAST search
3. Target protein was not hit by MS BLAST search
1. Expected peptide as confident hit
2. Expected peptide as low condident hit
Spectrum with lowest S/N
Spectrum with lowest S/N
PepNovo score
MASCOT score
PepNovo score
MASCOT score

Figure 2.29: Workflow representing *in silico* simulation of signal-to-noise ratio of peptide spectra for evaluation of the PepNovo/MS BLAST potential to positively validate the assignment of spectra.

The dataset contained 100 high-quality peptide spectra, which confidently matched upon MASCOT search a peptide sequence of in average 12 amino acid residues with ions scores > 70. A dedicated script (written by Henrik Thomas, Shevchenko Group, MPI-CBG) reduced the absolute intensities of all peaks with relative intensities above 1% of the base peak intensity with the steps of 1% and produced the series of 100 spectra with gradually altered signal-to-noise ratios. Their *dta* files were merged into a single *mgf* file and submitted to MASCOT search, and ions scores of spectra matched to the correct database peptide sequences were registered. The same *mgf* file was sequenced *de novo* by the PepNovo program in a batch mode, recording up to seven sequence candidates for each interpreted spectrum. PepNovo scores of predicted sequences were registered, and sequences were merged into a query and submitted to MS BLAST searches. The outcome was sorted in three groups: 1) where MS BLAST produced a hit that was confident according to MS BLAST scoring scheme (first group); 2) where the target peptide was listed in the output of the MS BLAST search as a borderline or nonconfident hit (second group); 3) or where the target protein was not hit by MS BLAST at all (third group). In the first and the second groups two spectra were identified (if possible) with the lowest signal-to-noise ratio and their ion scores (MASCOT), sequence quality scores, and MS BLAST scores were registered.



```
Monoisotopic mass of neutral peptide Mr(calc): 1528.70
Variable modifications:
M11     : Oxidation (M)
Ions Score: 84  Expect: 8.9e-07
Matches (Bold Red): 22/118 fragment ions using 30 most intense peaks
```

de novo sequences (PepNovo)

PepNovo score 11.3

BNQVYSAEDLEM.SK
BNQVYSAEDLEFSK
BNGAVYSAEDLEM.SK
BNGAVYSAEDLEFSK
BGGQVYSAEDLEM.SK
BGGGAVYSAEDLEM.SK
BGGQVYSAEDLEFSK

MS BLAST search — Snf5 - Caenorhabditis elegans

Score = 82 (42.2 bits)
Identities = 11/11 (100%), Positives = 11/11 (100%)

```
Query:     4 VYSAEDLEMSK 14
             VYSAEDLEMSK
Sbjct:   159 VYSAEDLEMSK 169
```

```
Monoisotopic mass of neutral peptide Mr(calc): 1528.70
Variable modifications:
M11     : Oxidation (M)
Ions Score: 37  Expect: 0.038
Matches (Bold Red): 12/118 fragment ions using 29 most intense peaks
```

de novo sequences (PepNovo)

PepNovo score 6.8

BNEVYSAEXXXXXEFSK
BNEVYSAEXXXXEFSK
BNEVYSAEXXXEFSK
BNEVYSAEXXXXXEM.SK
BNEVYSAEXXXEM.SK
BNEVYSAEXXEM.SK
BGGEVYSAEXXXXXEFSK

MS BLAST search — Snf5 - Caenorhabditis elegans

Score = 69 (35.0 bits)
Identities = 9/11 (81%), Positives = 9/11 (81%)

```
Query:    85 VYSAEXXEMSK 95
             VYSAE  EMSK
Sbjct:   159 VYSAEDLEMSK 169
```

Figure 2.30: Altering MS/MS spectra for in *in silico* simulation experiments.

(a) The presented tandem mass spectrum assigned upon MASCOT search the peptide (K)ELVYSAEDLEMSK from *C. elegans* protein Snf5 with ions score of 84; (c) a spectrum with altered signal-to-noise ratio, produced from the spectrum in panel (a) by reducing the intensity of fragment ions by 95%, while maintaining the same intensity of noise peaks. MASCOT search identified the same peptide, albeit the ions score was 37. (b) *De novo* interpretation of the spectrum in panel (a) by PepNovo software produced seven partially redundant candidate sequences, with the top candidate having a quality score of 11.3. According to MS BLAST conventions, (M.) stands for mono-oxidized methionine residues, and B stands for a generic trypsin cleavage site (arginine or lysine residues) preceding the peptide sequence. Since isobaric oxidized methionine and phenylalanine residues were not distinguished in ion trap spectra, both candidate sequences were included into the query string for MS BLAST search, which also produced a confident hit (MS BLAST confidence threshold score for a single reported high scoring pair (HSP) was 64). (d) The same procedure was applied to the modified spectrum from panel (c). Both ions score and PepNovo score decreased, yet MS BLAST search was still able to produce a confident hit.

2.2.2 There is a correlation between MASCOT ions scores and PepNovo quality scores?

I first checked if MASCOT ions scores and PepNovo quality scores correlated when both interpretations of the same marginal quality spectrum pointed to the same correct peptide sequence (Figure 2.31).

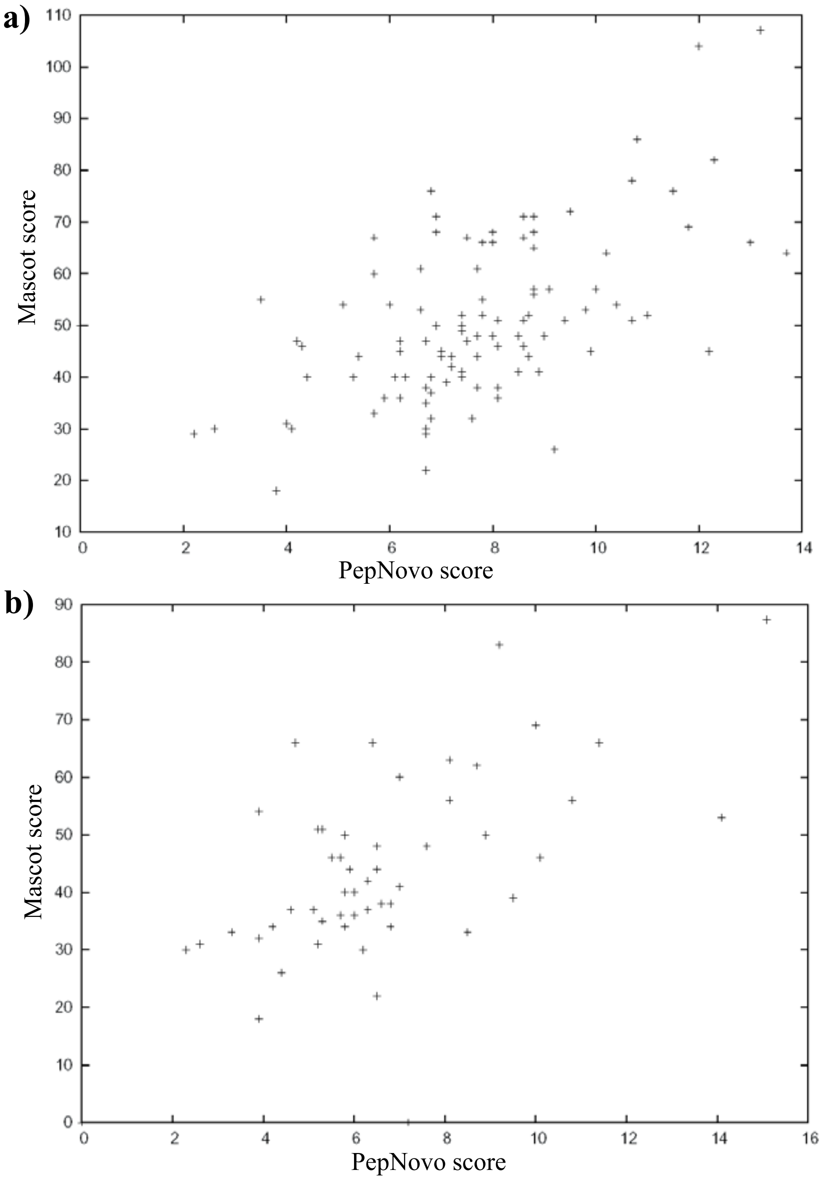


Figure 2.31: Plotted diagram of MASCOT ions scores versus PepNovo sequence quality scores.

Diagrams are built using the series of simulated spectra (Figure 2.29) that enabled their confident (panel a, data for 94 spectra) and nonconfident (panel b, data for 48 spectra) assignment to the correct database sequences by MS BLAST.

Although weak correlation was observed, I noticed that PepNovo scores corresponding to spectra with a given MASCOT score (or vice versa) varied within a broad range of values. This indicated that the two interpretations were, indeed, complementary and in many instances could independently cross-validate each other [3].

2.2.3 Validation of MS/MS spectra assignment

Figure 2.32 presents cumulative distributions of PepNovo scores (panel a) and MASCOT ion scores (panel b) obtained for the same dataset of *in silico* modified peptide spectra (Figure 2.29). They provide a complementary view on the ability of MS BLAST (Figure 2.32a) and PepNovo/MS BLAST combination (Figure 2.32b) to positively validate the assignment of spectra, depending on their PepNovo scores and MASCOT ions scores, respectively.

More than 60% of spectra, in which candidate peptide sequences were produced with PepNovo scores above 8, were confidently (according to MS BLAST scoring) matched to the correct protein entries by MS BLAST (Figure 2.32a), and for almost 80% of these spectra, correct peptide sequences were listed in search outputs. Once PepNovo scores exceeded 10, more than 90% of these spectra were confidently matched. This provided us with a qualitative estimate of the *de novo* interpretation reliability, irrespective of the actual MASCOT ions scores of examined spectra.

Using the same spectra dataset, we plotted the cumulated proportion of positive PepNovo/MS BLAST assignments of spectra against their MASCOT ions scores (Figure 2.32b). It should be noted that ions scores do not depend on the database size in contrast to thresholds scores of statistical confidence for performed MASCOT searches (Figure 2.32b).

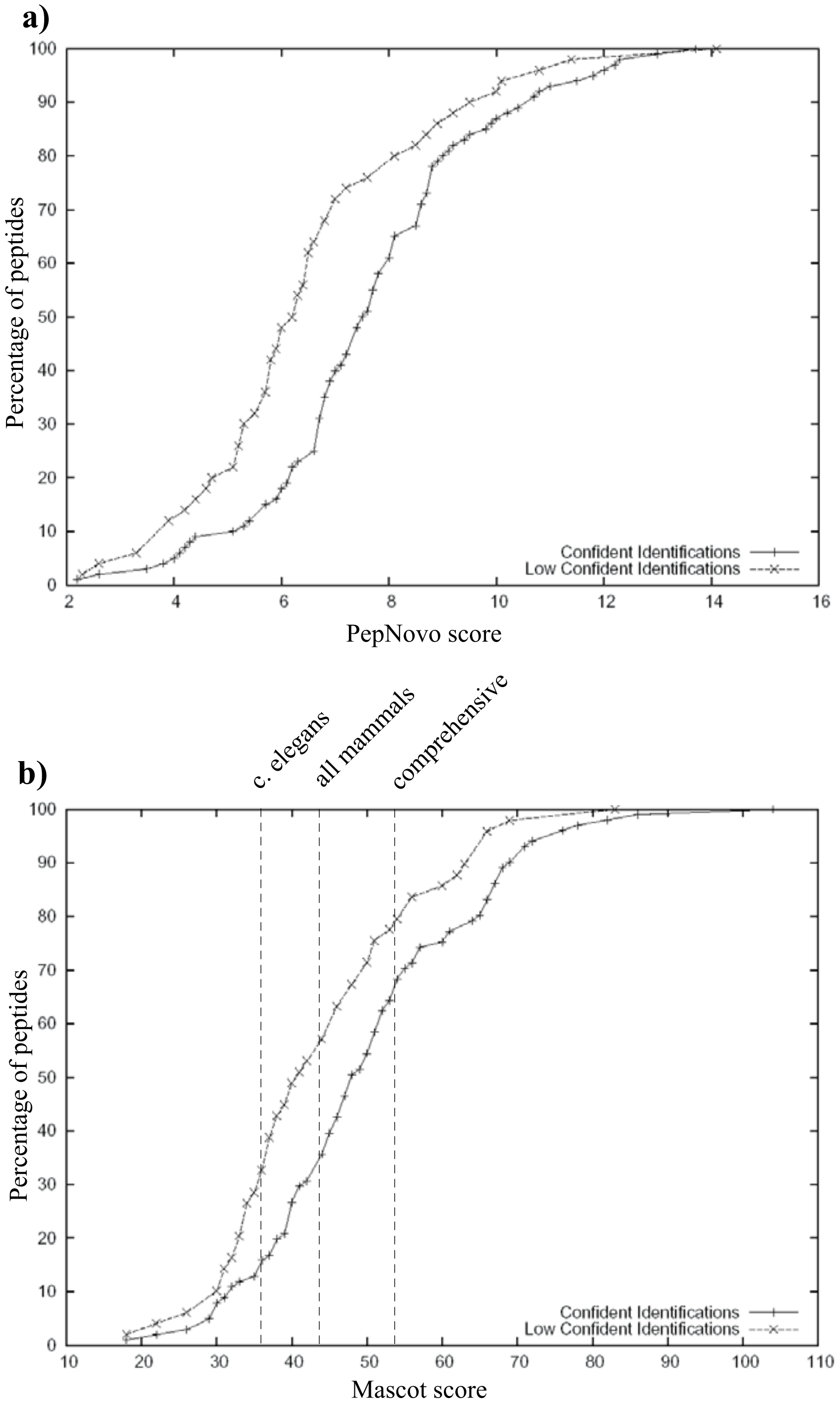


Figure 2.32: Cumulative distributions of confident and low confident MS BLAST hits obtained by searches with *de novo* sequences produced from tandem mass spectra with altered signal-to-noise ratio are plotted against their PepNovo scores (panel a) and MASCOT ions scores (panel b).

The dataset was the same as in Figure 2.31. Vertical bars in panel b stand for MASCOT thresholds of statistically confident protein identifications supported by matching a single peptide ($p < 0.05$) in the organism-specific databases: *C. elegans* (30 304 proteins entries), threshold score of 36; all mammals (287 223 protein entries), threshold score of 43; a comprehensive (all species) database (2 011 425 protein entries), threshold score of 53.

To positively identify a protein in a comprehensive (all species) database, the ions score of a one peptide hit should exceed a relatively high threshold (>53), even at the moderate $p < 0.05$. Therefore, positive protein identifications with one or two matched peptides would require exceptional quality of corresponding MS/MS spectra, and therefore, false negatives are common. For searches in smaller, species-restricted databases, threshold scores are lower (Figure 2.32b). These searches, however, often produce false positives by matching the spectra of peptides from exogenous protein contaminants to sequences of the assumed organism.

Figure 2.32b suggests that approximately 80% of borderline (potentially, false negative) one-peptide hits produced by searches against a comprehensive database should be directly verifiable via *de novo* sequencing and MS BLAST. Although the expected success rate also remains substantial for smaller species-restricted databases, *de novo* verification would be most helpful in discriminating against false positive, rather than validating false negative hits. Ions scores of false-positive hits are often marginal, since they are falsely matched to wrong database entries, although rich patterns of fragment ions together with low chemical noise enable confident readout of long stretches of their sequences [3].

2.2.4 The protein identification and validation workflow

A protein identification and validation routine employed in my work is depicted in Figure 2.33 and started with the stringent database search against a species-restricted database, in order to minimize the analysis time and to identify low abundant proteins whose spectra represent limited information (less than 3 peptides, poor quality, noisy spectra) [3].

It should be noted that different proteomics laboratories apply varying confidence criteria, even if the same software was used for database mining [188, 302]. The database independent validation by PepNovo/MS BLAST allowed me to use conserved criteria of positive protein identifications together with relatively loose selection of nonconfident hits, although this strategy yielded a large number of borderline hits. Many of them were produced by matching one or two spectra, and therefore, we could use their ions scores as direct selection criteria (Figue 2.33).

Since background proteins increase number of false positives by search in species restricted database first step in validation of borderline hits was their conformation in a comprehensive database. To this end corresponding dta files were fetched by Windows Shell Scripts developed in-house (Henrik Thomas, Shevchenko Group, MPI-CBG) and re-submitted to another round of MASCOT searches, now against a full database with unrestricted species specificity. The second search typically identified and removed good quality spectra of full tryptic peptides, originating from trypsin, keratin, GST, and other background proteins, which produced statistically confident hits in searches against a full database.

The remaining spectra were interpreted *de novo*, and the obtained sequence candidates were merged into a single query [224, 225] and searched against a comprehensive database by MS BLAST. The results of MS BLAST searches were interpreted as follows: if the same peptide as in the MASCOT search was either confidently matched by MS BLAST, or was present in the output of the MS BLAST search, and the reported high-scoring segment pair [303] (HSP), which corresponds to the alignment of the database peptide sequence and the sequence deduced from MS/MS spectrum by its *de novo* interpretation [225], covered at least 50% of the verified peptide sequence, then these hits were considered confirmed. It should be noted, that by PepNovo produced sequence proposals were not fully accurate, especially in case of poor quality and noisy target spectra. The predicted *de novo* sequences might contain correct sequence stretches that, however, did not produce statistically significant alignments and therefore were not reported within an HSP. In most cases, the length of the aligned non-interrupted peptide sequences exceeded six amino acid residues.

The MASCOT hits were considered as false positives and rejected if MS BLAST searches either confidently hit another protein, or hit a common background protein and more than 50% of the peptide sequence (and, at least, 6 amino acid residues) were covered by the aligned HSP.

The third criterion came from the consideration of the expected *de novo* interpretation accuracy, which is related to the PepNovo quality score. If *de novo* interpretation of validated MS/MS spectra produces peptide sequence candidates with PepNovo scores above 10, then, according to Figure 2.32, it was expected that subsequent MS BLAST searches would confirm more than 90% of the corresponding MASCOT hits. Otherwise, these hits were considered false positives, even if MS BLAST searches produced no significant alignments to other proteins.

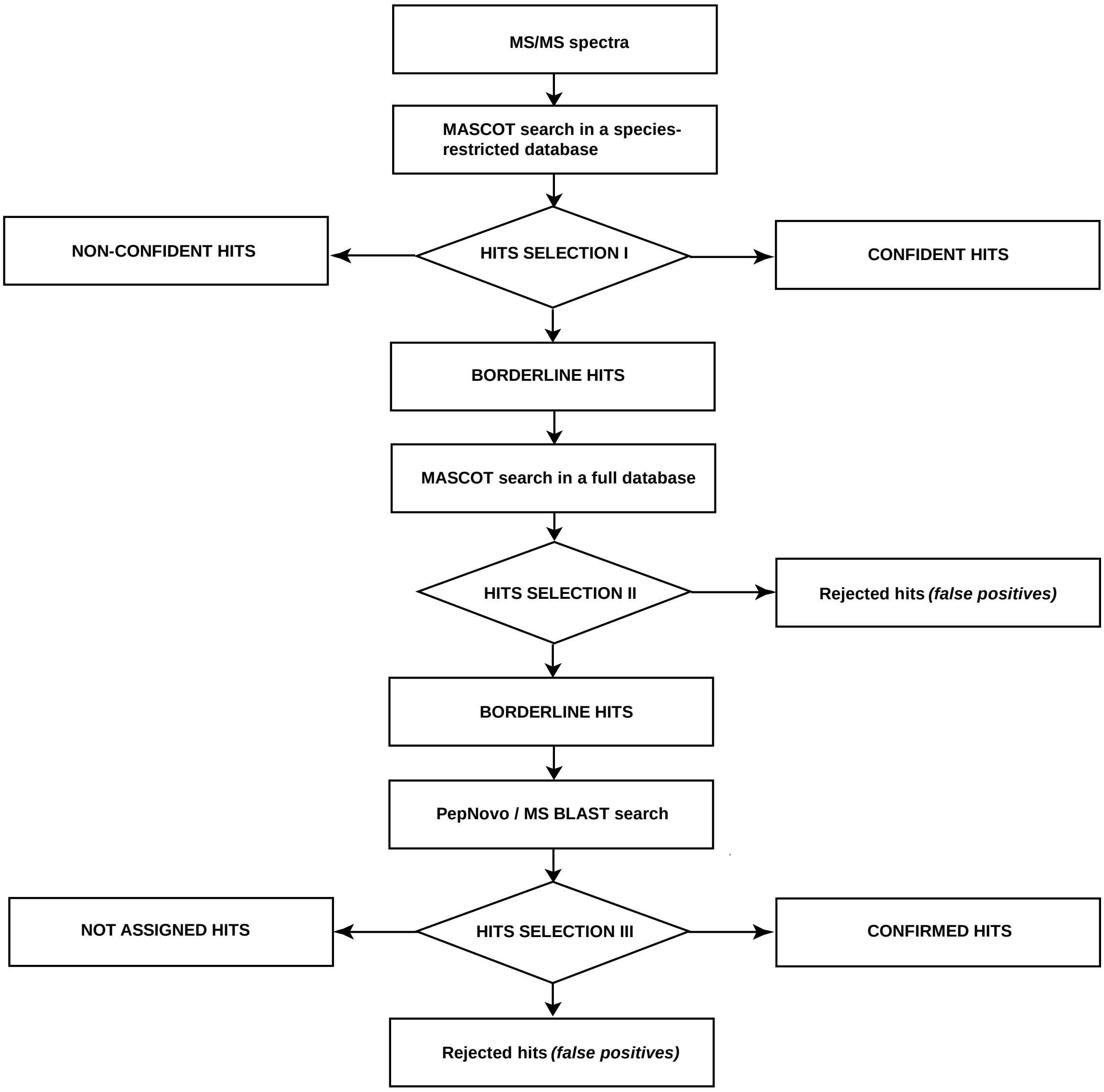

MS/MS spectra
MASCOT search in a species-restricted database
NON-CONFIDENT HITS
HITS SELECTION I
CONFIDENT HITS
BORDERLINE HITS
MASCOT search in a full database
HITS SELECTION II
Rejected hits (false positives)
BORDERLINE HITS
PepNovo / MS BLAST search
NOT ASSIGNED HITS
HITS SELECTION III
CONFIRMED HITS
Rejected hits (false positives)

Figure 2.33: Protein identification workflow that involves PepNov/MS BLAST vaidation of borderline hits.

Diamonds stand for the workflow junctions, where the following selection criteria were applied. Hit selection I: (1) confident hits: more than three peptides matched by MASCOT with ions scores above the confidence threshold for a species-specific database (36 for *C. elegans* protein database), or at least one score was above the threshold for a comprehensive database (53 for MSDB). (2) Borderline hits: MASCOT matched less than four peptides and the ions score of at least one peptide was within the range of ±30% of the threshold score (from 26 to 46 for *C. elegans*). (3) Nonconfident hits: the rest. Hit selection II: (1) rejected hits: the searched peptide confidently hit other than expected protein in a comprehensive database (ions score should exceed 53). (2) Borderline hits: the rest. Hit selection III: (1) confirmed hits: hits either confidently matching the expected protein by MS BLAST or in which the aligned HSP covered more than 50% of the expected peptide sequence spanning over more than six amino acid residues. (2) Rejected hits: common background proteins (trypsin, keratins, GST) matching the same criteria; or other proteins confidently matched by MS BLAST; or hits that did not match the expected peptide albeit their PepNovo scores were above 10. (3) Not assigned hits: the rest.

Low PepNovo scores (practically, less than 5) indicated that, for any reason, PepNovo failed to produce a reliable sequence of sufficient length. In these cases, negative outcomes of MS BLAST searches were inconclusive and the hits remained unassigned.

2.2.5 False positives and false negative hits revealed by PepNovo/MS Blast: case studies

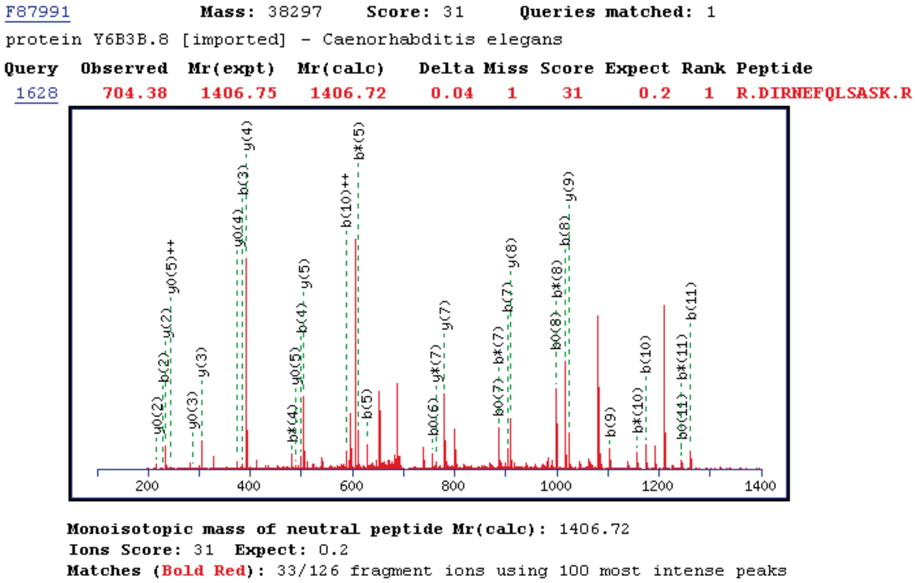
To demonstrate the practical applicability of the proposed workflow, here is presented the validation of two borderline hits produced by nanoLC-MS/MS analysis of gel-separated *C. elegans* proteins [4].

The protein Y6B3B.8 was identified by MASCOT search in a *C. elegans* protein database under the fixed trypsin cleavage specificity settings. The protein was hit by a single MS/MS spectrum with the ions score of 31 (Figure 2.34), while the proposed confidence threshold for *C. elegans* database was 36. Manual inspection of the spectrum suggested that almost all abundant peaks matched m/z of expected fragment ions.

To validate this hit, the corresponding spectrum was first searched against a comprehensive database. The search pointed to the same protein; however, the confidence of the identification was low, because of the increased database size. Therefore, the hit was further validated by PepNovo/MS BLAST (Figure 2.34b), which confidently hit a half-tryptic peptide VVEGNEQFISASK that originated from bovine trypsin, presumably via orifice fragmentation of the abundant autodigestion product LDEDNINVVEGNEQFISASK. It should be noted that approximately the same number of peaks matched the expected fragment ions in panels (a) and (c) of Figure 2.34, illustrating that manual inspection might be biased.

To further check the MS BLAST identification, another MASCOT search was performed without restricting the enzyme cleavage specificity. The search against a full species database resulted in the same trypsin peptide identified by PepNovo/MS BLAST. Despite higher ions score (67 for trypsin peptide versus 31 for *C. elegans* peptide), the hit was still nonconfident since the threshold score under the assumed settings was 74. The Expect value (the expected number of false-positive hits produced by searching a database with the given spectrum) was not improved and stayed well within the nonconfident range.

a)



```
F87991              Mass: 38297    Score: 31     Queries matched: 1
protein Y6B3B.8 [imported] - Caenorhabditis elegans
Query  Observed  Mr(expt)  Mr(calc)    Delta Miss Score Expect Rank Peptide
 1628    704.38   1406.75   1406.72     0.04   1     31     0.2    1  R.DIRNEFQLSASK.R
```

```
Monoisotopic mass of neutral peptide Mr(calc): 1406.72
Ions Score: 31   Expect: 0.2
Matches (Bold Red): 33/126 fragment ions using 100 most intense peaks
```

b) de novo sequences (PepNovo)

PepNovo score 11.4

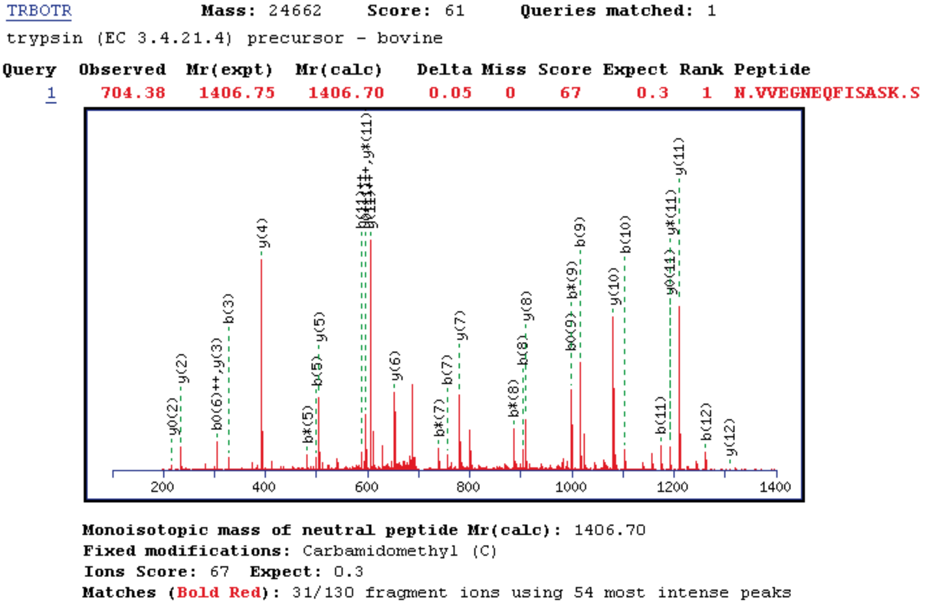
BVVEGNEQM.LSASK
BVVEGNEQFLSASK
BVVEGNEGAM.LSASK
BVVEGNEGAFLSASK
BVVEGGGEQM.LSASK
BVVEGGGEQFLSASK
BVVEGGGEGAM.LSASK

MS BLAST search   trypsin, bovine

Score = 91 (47.6 bits)
Identities = 12/13 (92%), Positives = 13/13 (100%)

```
Query:     17 VVEGNEQFLSASK 29
              VVEGNEQF+SASK
Sbjct:     77 VVEGNEQFISASK 89
```

c)



```
TRBOTR              Mass: 24662    Score: 61     Queries matched: 1
trypsin (EC 3.4.21.4) precursor - bovine
Query  Observed  Mr(expt)  Mr(calc)    Delta Miss Score Expect Rank Peptide
   1     704.38   1406.75   1406.70     0.05   0     67     0.3    1  N.VVEGNEQFISASK.S
```

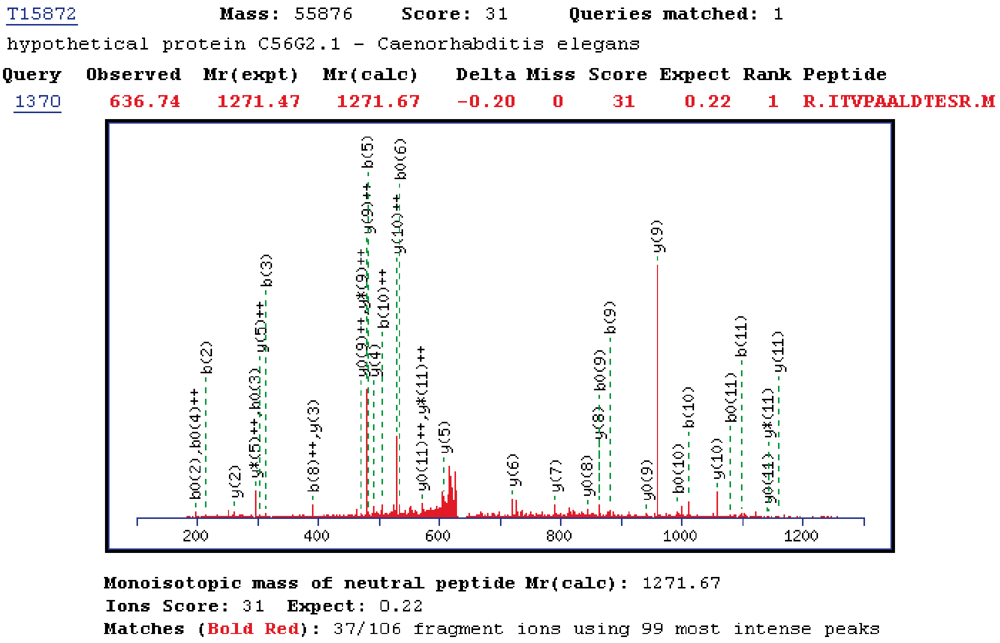
```
Monoisotopic mass of neutral peptide Mr(calc): 1406.70
Fixed modifications: Carbamidomethyl (C)
Ions Score: 67   Expect: 0.3
Matches (Bold Red): 31/130 fragment ions using 54 most intense peaks
```

Figure 2.34: PepNovo/MS BLAST validation of the protein identification with borderline statistical confidence: example of a false-positive hit.

(a) MASCOT search performed against the *C. elegans* database hit the protein Y6B3B.8. Search against a full database also confirmed this hit. Trypsin was specified as proteolytic enzyme in both searches. (b) The same spectrum was interpreted *de novo*, and candidate sequences were submitted to MS BLAST search, which matched the half-tryptic peptide VVEGNEQFISASK from bovine trypsin as a single confident hit. (c) The same spectrum as in panel (a) with fragment ions matching the sequence of VVEGNEQFISASK. MASCOT search was here performed without restricting the enzyme cleavage specificity in a comprehensive database; the threshold score under the assumed settings was 74.

In another case, *C. elegans* protein C56G2.1 was identified by matching one peptide with ion score of 31, which is below the threshold ion score for the chosen database (Figure 2.35). MASCOT search against a comprehensive database with and without trypsin cleavage specificity restrictions also pointed to the same protein, although both ions scores were statistically insignificant. At the same time, *de novo* interpretation of the spectrum followed by MS BLAST search confidently hit the expected peptide sequence from C56G2.1 protein, thus, rescuing this, otherwise false negative, hit (Figure 2.35b).

a)



```
T15872              Mass: 55876     Score: 31      Queries matched: 1
hypothetical protein C56G2.1 - Caenorhabditis elegans
Query  Observed  Mr(expt)  Mr(calc)   Delta Miss  Score Expect Rank Peptide
 1370    636.74   1271.47   1271.67   -0.20    0     31    0.22    1  R.ITVPAALDTESR.M
```

```
Monoisotopic mass of neutral peptide Mr(calc): 1271.67
Ions Score: 31  Expect: 0.22
Matches (Bold Red): 37/106 fragment ions using 99 most intense peaks
```

b)

de novo sequences
(PepNovo)

PepNovo score 5.8

BLTVPAALDTESR
BTLVPAALDTESR
BXXVPAALDTESR
VPAALDTESR
VPAALDTE
BLTVPAAL
BTLVPAAL

MS BLAST search   C56G2.1 - Caenorhabditis elegans

Score = 86 (44.7 bits)
Identities = 11/13 (84%), Positives = 13/13 (100%)

```
Query:     1 BLTVPAALDTESR 13
             ++TVPAALDTESR
Sbjct:   838 RITVPAALDTESR 850
```

Figure 2.35: PepNovo/MS BLAST validation of the protein identification with borderline statistical confidence: example of a false-negative hit.

(a) MASCOT search against *C. elegans* database hit C56G2.1 protein with insignificant ions score. Search against a comprehensive database pointed to the same protein. (b) The same spectrum was interpreted *de novo*, and candidate sequences were searched by MS BLAST that confidently hit the same *C. elegans* protein.

2.2.6 Validating of borderline hits at the large scale: biological applications

2.2.6.1 Determination of interaction partners of the protein TPXL-1 required for mitotic spindle assembly in C. elegans.

Functional analysis of an uncharacterized novel gene *tplx-1*, performed by Nurhan Özlü (Tony Hyman, MPI-CBG) revealed that TPLX-1 is the invertebrate orthologue of TPX2, which is known from studies in *Xenopus* and mammalian cells to be involved in mitotic spindle assembly, and to interact with Aurora A kinases [304].

A genome-wide Yeast Two-Hybrid screen of *C.elegans* proteins identified an interaction between TPXL-1 and AIR-1 [305]. To test if TPXL-1 and AIR-1 form a complex *in vivo* and determine other possible interaction partners, Nurhan Özlü performed a GST pull-down using GST::TPXL-1 as a bait. The eluted proteins were separated by SDS-PAGE and subjected to nano-LC-MS/MS analysis.

In the analysis of 10 Coomassie-stained bands, 127 proteins (44%) were confidently identified, among them AIR-1, confirming the assumption that both protein interact, and another 164 hits (56%) were regarded borderline (Figure 2.36), according to the criteria discussed above (Figure 2.33).

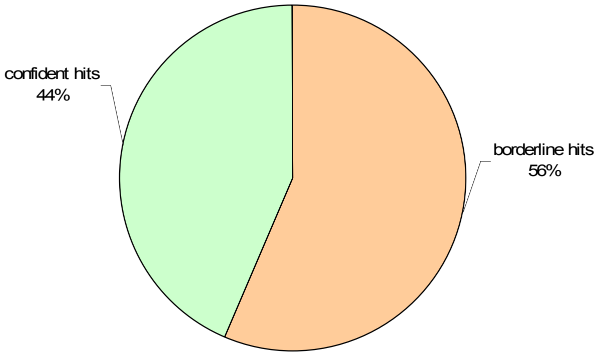
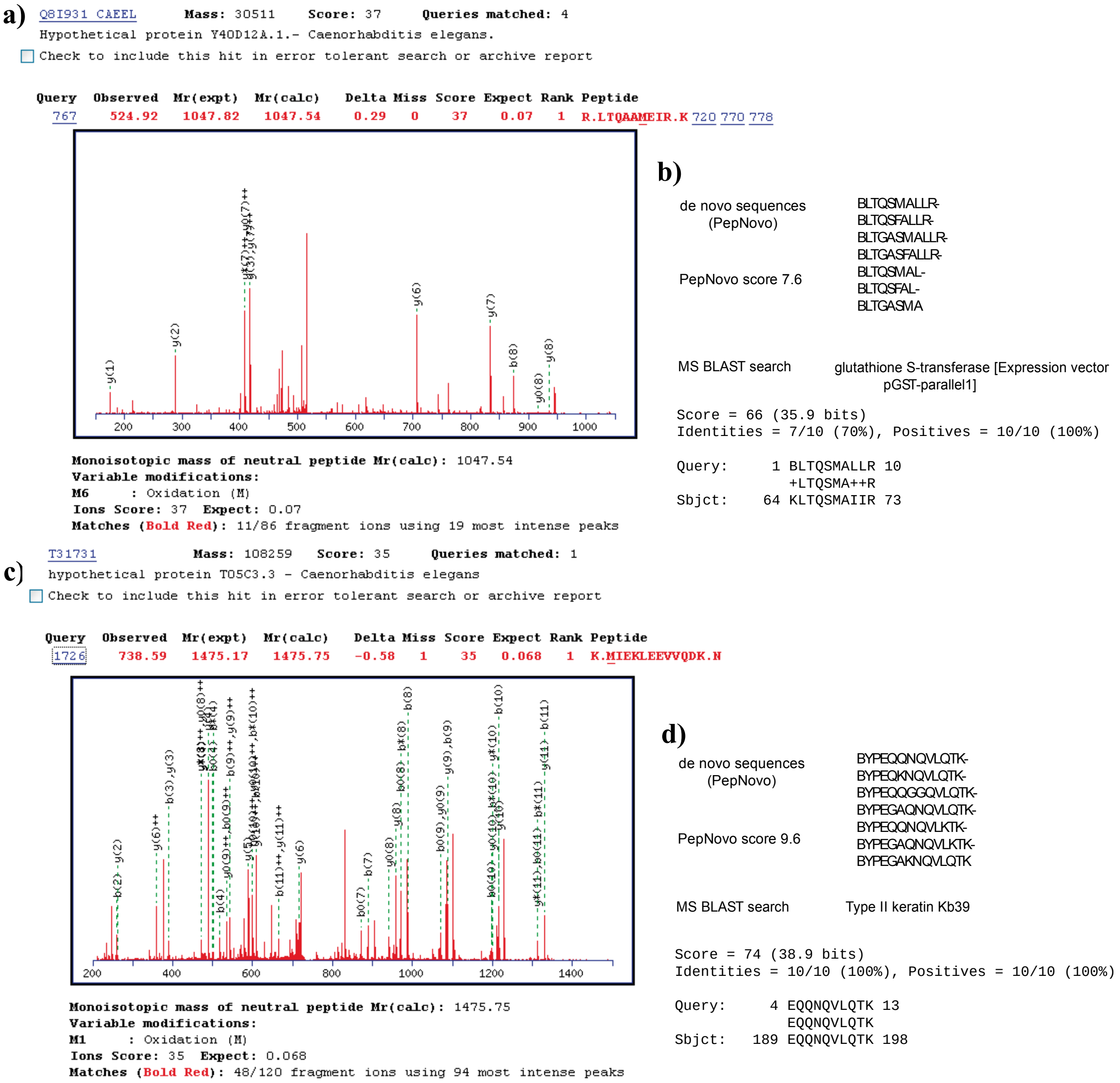


Figure 2.36: Fraction of borderline hits obtained by LC-MS/MS analysis of the GST pull-down experiment performed to study protein-protein interactions of the *C.elegens* protein TPLX-1.

GST pull-down was performed using GST::TPLX-1 expressed in *E.coli*. GST::TPLX-1 was bound to glutathione beads; the prepared worm extract was incubated with the resin, washed and the bound proteins were eluted by adding reduced glutathione. The eluted proteins were separated by SDS-PAGE and stained with Coomasie (Nurhan Özlu, MPI-CBG). 10 bands were cut and in-gel digested; subsequently the obtained peptides were extracted from the gel matrix and analysed by nano-LC-MS/MS. The acquired tandem mass spectra were searched by MASCOT against the *C. elegans* database. 127 proteins were confidently identified and another 164 hits were regarded borderline.

Many of these hits were of substantial biological interest. However, searching MS/MS spectra against a comprehensive database revealed that the preparation was heavily contaminated with exogenous proteins, such as fragments of the GST construct, proteins from *Escherichia coli* (host organism in which the GST-fused bait protein was expressed), and human keratins (Figure 2.37).



a)

```
Q8I931_CAEEL       Mass: 30511    Score: 37     Queries matched: 4
Hypothetical protein Y40D12A.1.- Caenorhabditis elegans.
□ Check to include this hit in error tolerant search or archive report

Query  Observed  Mr(expt)  Mr(calc)   Delta Miss Score Expect Rank Peptide
  767    524.92   1047.82   1047.54    0.29    0    37    0.07    1  R.LTQAAMEIR.K 720 770 778
```

```
Monoisotopic mass of neutral peptide Mr(calc): 1047.54
Variable modifications:
M6      : Oxidation (M)
Ions Score: 37  Expect: 0.07
Matches (Bold Red): 11/86 fragment ions using 19 most intense peaks
```

b)

de novo sequences (PepNovo)

PepNovo score 7.6

BLTQSMALLR-
BLTQSFALLR-
BLTGASMALLR-
BLTGASFALLR-
BLTQSMAL-
BLTQSFAL-
BLTGASMA

MS BLAST search   glutathione S-transferase [Expression vector pGST-parallel1]

```
Score = 66 (35.9 bits)
Identities = 7/10 (70%), Positives = 10/10 (100%)

Query:     1 BLTQSMALLR 10
               +LTQSMA++R
Sbjct:    64 KLTQSMAIIR 73
```

c)

```
T31731              Mass: 108259    Score: 35      Queries matched: 1
hypothetical protein T05C3.3 - Caenorhabditis elegans
□ Check to include this hit in error tolerant search or archive report

Query  Observed  Mr(expt)  Mr(calc)   Delta Miss Score Expect Rank Peptide
 1726    738.59   1475.17   1475.75   -0.58   1    35   0.068   1  K.MIEKLEEVVQDK.N
```

```
Monoisotopic mass of neutral peptide Mr(calc): 1475.75
Variable modifications:
M1      : Oxidation (M)
Ions Score: 35  Expect: 0.068
Matches (Bold Red): 48/120 fragment ions using 94 most intense peaks
```

d)

de novo sequences (PepNovo)

PepNovo score 9.6

BYPEQQNQVLQTK-
BYPEQKNQVLQTK-
BYPEQQGGQVLQTK-
BYPEGAQNQVLQTK-
BYPEQQNQVLKTK-
BYPEGAQNQVLKTK-
BYPEGAKNQVLQTK

MS BLAST search   Type II keratin Kb39

```
Score = 74 (38.9 bits)
Identities = 10/10 (100%), Positives = 10/10 (100%)

Query:     4 EQQNQVLQTK 13
               EQQNQVLQTK
Sbjct:   189 EQQNQVLQTK 198
```

Figure 2.37: Revealing of false positives by PepNovo/MS BLAST from the data obtained by nanoLC-MS/MS analysis of the GST pull-down experiment.

(a) MASCOT search performed against the *C. elegans* database hit the protein Y40D12A.1;
(b) The same spectrum was interpreted *de novo*, and candidate sequences were submitted to MS BLAST search, which matched tryptic peptide originated from GST construct; (c) Protein T05C3.3 was identified by MASCOT search against the *C. elegans* database. (d) MS BLAST search performed upon *de novo* predicted sequences of the corresponding MS/MS spectrum matched keratin as confident hit. Both spectra (a) and (c) were searched against the comprehensive database and matched GST and keratin peptides, respectively.

It should be noted that it is absolutely impractical to perform database searching against a comprehensive database, since it considerably increases analysis time. In addition all database searching algorithms rank identified hits according their scores, which basically express their abundance. Thus, low abundant proteins from the organism of interest might be ranked far below high abundant and totally irrelevant background proteins (in the presented case GST construct, keratins, *E. coli* proteins, bovine trypsin), significantly complicating data analysis.

My next intention, therefore, was to see how successful is database independent validation by PepNovo/MS BLAST in revealing false positives compared to the database searching in the comprehensive database. Since identification and validation of proteins was performed according to the workflow depicted in Figure 2.33 I could evaluate this.

Figure 2.38 presents a distribution of 164 validated borderline hits: 37% of them were confirmed by PepNovo/MS BLAST, whereas another 34% were discarded as false positives (either by PepNovo/MS BLAST or by MASCOT searches against a nonrestricted database), so that the percentage of borderline identifications that still remained ambiguous was reduced down to 29%.

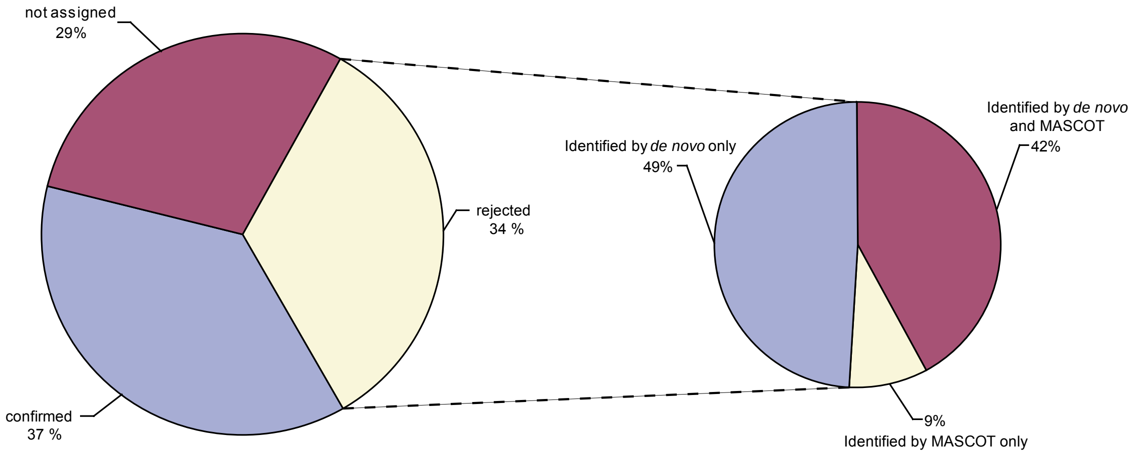


Figure 2.38: Validation of the borderline hits from the data obtained by nanoLC-MS/MS analysis of the GST pull-down experiment.

Confirmed: hits confirmed by PepNovo/MS BLAST method. "Rejected": hits were rejected if either MASCOT confidently identified another protein in a full (all species) database (designated as "Identified by MASCOT" at the inset), or by PepNovo/MS BLAST probing according to the workflow in Figure 2.33 (designated as "Identified by de novo"). "Not assigned": borderline hits for which both methods did not produce any conclusive identity evidence.

Interestingly, among recognized false positives, 49% were identified only by PepNovo/MS BLAST, 42% were verifiable by MASCOT searches against a full database as well as by PepNovo/MS BLAST, and only 9% were identified by MASCOT searches in the full-species database, while PepNovo/MS BLAST failed to produce conclusive assignments. Thus, 116 borderline hits (71%) were confirmed or rejected, and the total number of ambiguous identifications was considerably reduced without any recourse to manual inspection of spectra.

Taken together, in the performed study nanoLC-MS/MS analysis and database searching enabled identification of about 300 proteins, among them 56% were of borderline statistical confidence. Supported by database independent PepNovo/MS BLAST validation I could considerably reduce number of these ambiguous hits. The study confirmed as expected that TPXL-1 and AIR-1 are interaction partners. In further experiments Nurhan Özlü showed that the essential function of TPXL-1 is to activate and localize Aurora A to the mitotic spindle assembly. This provided mechanistic insight into how the converted TPX2 protein family contributes to spindle assembly [4].

2.2.6.2 Determination of RSA-1 associated proteins required for mitotic spindle assembly in C. elegans.

In the course of a genome-wide screening, the uncharacterized gene RSA-1 (for regulator of spindle assembly 1) was remarked because its silencing resulted in a dramatic spindle assembly defect. Annelore Schlaitz (Prof. Tony Hyman, MPI-CBG, Dresden) studied RSA-1 (RNAi) phenotype and found out that is required for two separable centrosomal pathways in spindle formation: 1) the promotion of microtubule outgrowth from centrosomes in a process downstream of tubulin-mediated nucleation and 2) the stability of kinetochore microtubules.

RSA-1 (C25A1.9) encodes 404 amino-acid protein with sequence similarity to B-type regulatory subunits of Protein Phosphotase 2A (PP2A), most closely related to B’’ subunits of the TON2 subfamily. Interestingly, the *Arabidopsis thaliana* B’’ PP2A subunit TON2 has been implicated in aspects of microtubule cytoskeleton organization [306].

In order to see whether RSA-1 indeed functions as PP2A regulatory subunit and to find the interactions partner of RSA-1, co-immunoprecipitations experiments were carried out. First, the anti-RSA-1 antibody was used to immunoprecipitate RSA-1 and associated proteins from extracts of *C. elegans* embryos. NanoLC-MS/MS analysis of this IP resulted in a large number of proteins (including high number of borderline hits, which have been validated by PepNovo/MS BLAST), among them the core centrosomal protein SPD-5 and the uncharacterized protein Y48A6B.11. Interestingly, Y48A6B.11 had been previously found to interact directly with RSA-1 in a large-scale yeast-two-hybrid screen [305]. However, no phosphotase subunits were detected in this preparation.

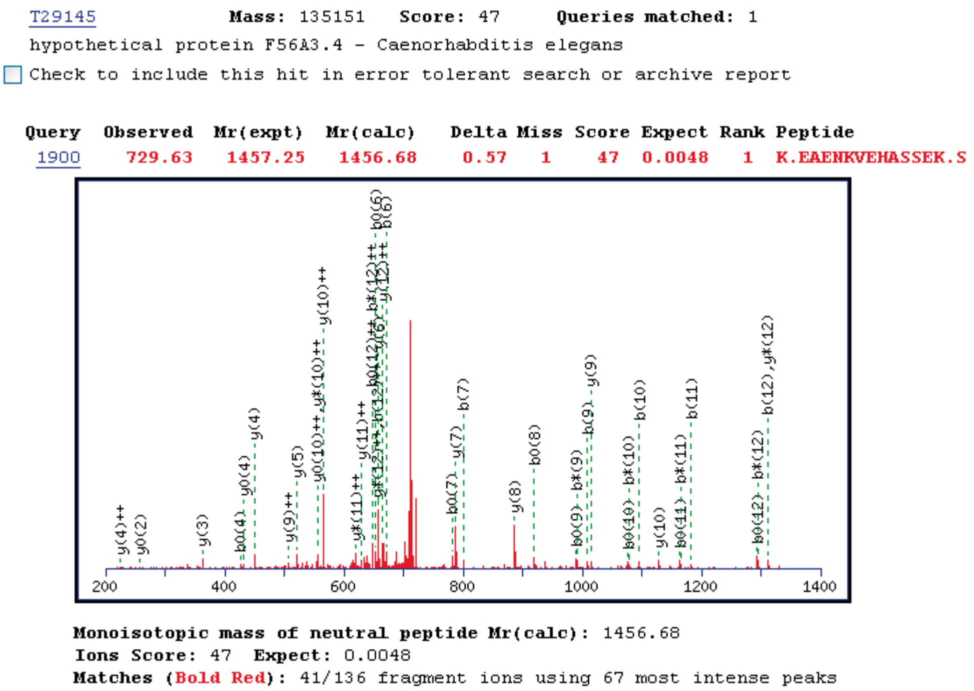
The antibody only binds the extreme C-terminus of RSA-1, which is the region with the largest sequence conservation among regulatory B-subunits. One explanation for failure to detect PP2A subunits might therefore be that the anti-RSA-1 antibody only precipitates the fraction of RSA-1 that is not engaged in a PP2A complex and that the epitope recognized by this antibody is not accessible in the heterotrimetric complex.

Annelore Schlaitz therefore chose a different approach for immunoprecipitating the protein, using a worm strain that expressed GFP-tagged RSA-1. Extracts were prepared from GFP::RSA-1 worms and subjected to co-immunoprecipitation with anti-GFP antibodies, followed by nano LC-MS/MS analysis. This IP experiment was performed twice, using different controls: in the first experiment, random IgG antibodies were incubated with GFP::RSA-1 extract; in the second experiment, the anti-GFP antibody was incubated with extracts from wild-type worms that did not express the transgene. Moreover, salt conditions were modified, as compared to the previous, high-background anti-RSA-1 IP, resulting in far fewer co-purifying proteins.

Although, the second IP experiment was more efficient, several proteins identified here were of borderline statistical confidence. Among them was protein SPD-5, which was expected to be associated with RSA-1. SPD-5 was hit by a single peptide (K)EAENKVEHASSEK upon MASCOT search in a *C. elegans* database (confidence threshold 36) with the ion score of 47 (Figure 2.39a). According to the selection criteria described before (chapter 2.2.4) this hit was considered as borderline and subjected to further validation. To this end the corresponding MS/MS spectrum was sequenced *de novo* (PepNovo), and predicted sequence candidates were submitted to MS BLAST search. MS BLAST could confidently confirm this hit (Figure 2.39b).

The results of both IP experiments were combined and only those proteins identified in both preparations were considered. Four proteins were found to reproducibly and specifically associate with RSA-1. These were the Protein Phosphatase 2A catalytic subunit LET-92 and the PP2A structural subunit PAA-1 as well as Y48A6B.11 and SPD-5, two proteins that had already been found through immunoprecipitations using the antibody against the endogenous protein (Figure 2.40).

a)



```
T29145              Mass: 135151    Score: 47     Queries matched: 1
hypothetical protein F56A3.4 - Caenorhabditis elegans
□ Check to include this hit in error tolerant search or archive report

Query  Observed  Mr(expt)  Mr(calc)   Delta Miss Score Expect Rank Peptide
 1900    729.63   1457.25   1456.68    0.57   1    47  0.0048   1  K.EAENKVEHASSEK.S
```

```
Monoisotopic mass of neutral peptide Mr(calc): 1456.68
Ions Score: 47  Expect: 0.0048
Matches (Bold Red): 41/136 fragment ions using 67 most intense peaks
```

b)

de novo sequences (PepNovo)

PepNovo score 8.5

BXXXXNQVEHASSEK-
BXXXXNGAVEHASSEK-
BXXXXGGQVEHASSEK-
BXXXXGGGAVEHASSEK-
NQVEHASSEK-
BXXXNQVEHASSEK-
BXXXNGAVEHASSEK

MS BLAST search (spd-5) Caenorhabditis elegans

```
Score = 67 (35.8 bits)
Identities = 9/14 (64%), Positives = 10/14 (71%)

Query:    80 BXXXNQVEHASSEK 93
               +    N VEHASSEK
Sbjct:   883 KEAENKVEHASSEK 896
```

Figure 2.39: PepNovo/MS BLAST validation of the *C. elegans* protein SPD-5.

(a) MASCOT search was performed against the *C. elegans* database and hit SPD-5. Search against a comprehensive database also confirmed this hit. Trypsin was specified as proteolytic enzyme in both searches. (b) The spectrum was interpreted *de novo*, and candidate sequences were submitted to MS BLAST search, which confidently confirmed SPD-5.

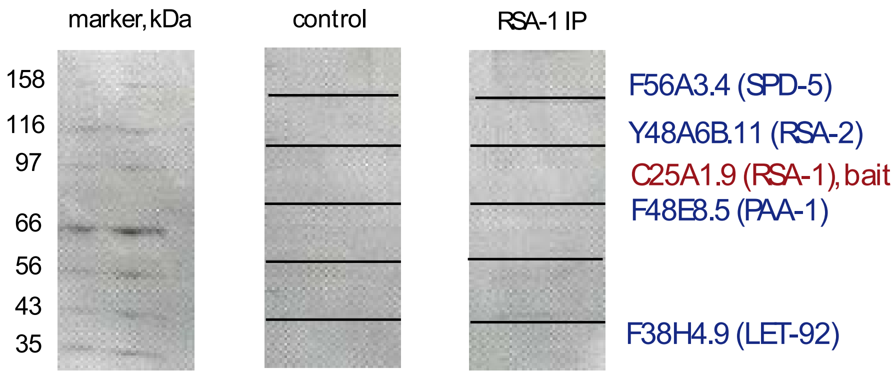


Figure 2.40: Proteins co-immunoprecipitating specifically with RSA-1.

Coomassie-stained gel represents protein marker (left panel), control (middle panel) and proteins, which specifically immunoprecipitate with RSA-1 (right panel).

Extracts were prepared from GFP::RSA-1 worms and subjected to co-immunoprecipitation with anti-GFP antibodies. The control for this experiment was prepared by incubating the anti-GFP antibody with extracts from wild-type worms that did not express the transgene. Gel lanes of control and immunoprecipitates were cut in 6 equal bands and analyzed by nano LC-MS/MS, followed by MASCOT database searching in the *C. elegans* database.

Taken together, supported by comprehensive LC-MS/MS analysis and PepNovo / MS BLAST validation we identified proteins associated with the novel protein RSA-1 (RSA complex). Further experiments allowed Annelore to discover and characterize a new regulatory pathway in *C. elegans* spindle assembly [5].

2.2.6.3 Validation by PepNovo/MS BLAST: what was achieved?

Performed studies demonstrated that combination of *de novo* sequencing by PepNovo and MS BLAST searches efficiently complements the conventional (based on database searching) protein identification routine. This method provides an independent means of automated validation of hits with borderline statistical confidence and substantially helps to reduce the rates of both false-positive and false-negative identifications.

3 CONCLUSION AND PERSPECTIVES

Bottom-up proteomics includes four important steps: 1) protein digestion, 2) peptide separation, 3) peptide fragmentation, and 4) data analysis.

Protein digestion is the most important step, in which proteins are cleaved in peptides of suitable size for mass spectrometric analysis. To address efficiency and completeness of in-gel digestion thermostable trypsin conjugates, obtained by modification of conventional bovine trypsin with oligosaccharides, were tested in accelerated in-gel digestion of proteins [2]. The modification of trypsin did not considerably increased its molecular weight (from ~25 to 33 kDa) but significantly improved its thermostability (for selected MAT-BT, RAF-BT, RAFR-BT and STA-BT, T_{50} increased by about 20°C) and suppressed autolysis, without affecting its cleavage specifity. MALDI TOF PMF of in-gel digests obtained by trypsin conjugates showed less autolytic peaks in the m/z range of 700 – 2700 compared to unmodified BT, simplifying protein identification.

To evaluate catalytic efficiency of trypsin conjugates in accelerated in-gel digestion of proteins a comprehensive kinetic study was carried out, where effect of the temperature, enzyme concentration and digestion time on the yield of digestion products was evaluated. To quantify in-gel digestion yield two different quantification approaches were tested and established: stable isotope labeling strategy, which employs ^{18}O-labeled peptide internal standards and is based on MALDI TOF analysis as well as label-free quantification approach, which utilizes mass spectral peak intensities of peptide ions from nanoLC-MS/MS data. Both quantification studies provided consistent results and demonstrated that the initially set goal to shorten sample preparation time and to improve recovery of conventional in-gel digestion is not realizable using glycosylated trypsins. Thus, at the best 60 to 70% of conventional digestion yield could be reached by in-gel digestion of proteins using trypsin conjugates (MAT-BT, RAF-BT and RAFR-BT) at accelerated conditions.

The obtained results suggested that one of the major factors responsible for the reduced in-gel digestion efficiency of the tested trypsin conjugates is their bulky structure (caused by the attached sugar chains), which significantly decreases their diffusion mobility in polyacrylamide gel matrix.

Therefore it seems to be reasonable to test glycosylated trypsins in gel-free shotgun proteomics, which relies on direct digestion of proteins in-solution. Current in-solution digestion of proteins is time-consuming and partially not efficient, especially when working with complex protein mixtures or hydrophobic and membrane proteins. Several additives such as surfactants, organic solvents, and urea are commonly used to denaturate proteins and to improve their solubility. However, such denaturants reduce the proteolytic activity of enzymes, setting an upper limit on applicable concentration and are often not compatible with mass spectrometry and liquid chromatography, requiring sample cleanup prior to LC or LC-MS/MS. The concentration limit of a denaturant is often below its desirable amount to fully denature and solubilise proteins in complex mixtures. This problem can be addressed by chemically modified trypsin conjugates, which have been reported to show noticeable resistance to denaturants such as urea, SDS and organic solvents [138, 141]. Therefore as next, the resistance of glycosylated trypsins to different denaturants should be tested in in-solution digestion of proteins at accelerated temperature, which also promotes denaturing conditions. These thermostable and autolysis resistant enzymes might find their use in analysis of complex protein mixtures, including hydrophobic proteins such as integral and transmembrane proteins, which have been a big challenge in proteomics since high concentrations of strong denaturants are required to solubilise them.

The next important proteomic problem addressed in the presented work concerns the reliability of protein identification based on database searching, pointing out problem of unrecognized false positives and borderline hits.

A validation method was developed and established [3], which employs database independent interpretation of the acquired tandem mass spectra by *de novo* sequencing software PepNovo combined with mass-spectrometry driven BLAST (MS BLAST) sequence similarity searching, which utilizes redundant, degenerate and partially accurate peptide sequence candidates and employs an independent scoring scheme to evaluate the confidence of database searching hits [220].

This validation approach was applied in a collaborating project, which aimed to prove *in vivo* interaction between *C. elegans* proteins TPXL-1 and AIR-1 and determine other possible interaction partners of the uncharacterized protein TPXL-1 [4]. NanoLC-MS/MS analysis of 10 in-gel digests of Coomassie-stained protein bands identified, in total, more than 290 proteins of varying abundance, among them 164 hits (56%) were of borderline confidence. Using a combination of MASCOT and PepNovo/MS BLAST searches, the assignment of more than 70% of borderline hits could be independently confirmed or rejected without manual inspection of raw MS/MS spectra. PepNovo/MS BLAST was further applied in another collaborating projects to validate borderline hits obtained by identification of proteins associated with the novel *C. elegans* protein RSA-1 (RSA complex) [5].

The presented study demonstrated that a combination of MASCOT software, *de novo* sequencing software PepNovo and MS BLAST, bundled by a simple scripted interface, enabled rapid and efficient validation of a large number of borderline hits, produced by matching of one or two MS/MS spectra with marginal statistical significance.

However, the method performance was inherently limited by the ability of *de novo* sequencing software to produce meaningful sequence candidates from tandem mass spectra with either insufficient fragment representation, or having too complex fragment patterns. Thus, it seems promising to employ simultaneously several independent peptide fragmentation methods within the same nanoLC-MS/MS experiment, which might increase the accuracy of *de novo* sequencing without compromising the analysis throughput and, presumably, sensitivity [180].

4 MATERIALS AND METHODS

4.1 Thermostable trypsin conjugates

4.1.1 Synthesis and bioanalytical characterization

Glycosylation of bovine trypsin and bioanalytical characterization of the obtained trypsin derivates was performed by Prof. Dr. Marek Šebela from Department of Biochemistry, Palacky University (Olomouc, Czech Republic) as described [1]. Trypsin glycosylation by disaccharides lactose, maltose and mellibiose was partially based on the protocol of Vaňková et al. [307]. Whereas modification of trypsin by trisaccharides maltotriose and raffinose, tetrasaccharide stachyose as well as α-/ β-cyclodextrines was based on the protocol of Morand and Biellmann [308]. Glycosylation of bovine trypsin was achieved by coupling oligosaccharides to its lysine residues. In addition, free argenyl residues in raffinose modified trypsin (RAF-BT) were optionally reacted with biacetyl [309], yielding an RAFR-BT with modified arginine residues.

Glycosylated enzymes were purified by ion exchange chromatography and dialyzed against 20 mM sodium acetate, pH 4.0 or 0.1% formic acid, concentrated by ultrafiltration, lyophilized and stored at –80 °C.

Trypsin activity was determined using a chromogenic substrate N$^{\alpha}$-benzoyl-DL-arginine-4-nitroanilide (BAPNA) as described [2].

Thermostability of trypsin and its conjugates was evaluated by monitoring the changes in their activity upon incubating enzyme aliquots in 20 mM sodium acetate buffer, pH 4.0 at 37 °C, 45°C, 55°C, 65°C and 75°C for 30 min.

Protein content was determined using a modified Lowry method [310].

Total carbohydrates were determined by the phenol-sulfuric acid method [311].

Primary amino groups were estimated by TNBS-reagent (2,4,6-trinitrobenzenesulfonic acid) [312].

For all obtained trypsin conjugates pI were determined according to the following protocol [313].

Molecular masses of modified trypsin conjugates were determined by tricine-SDS-PAGE according to Schägger et al.[314] and by MALDI-TOF MS.

4.1.2 Study of in-gel digestion kinetics using ^{18}O labeled peptides

4.1.2.1 Chemicals

All chemicals were purchased from Sigma-Aldrich (Steinheim, Germany) and were of analytical grade, unless otherwise noted. Concentrations of stock solutions of the standard proteins BSA, Aldolase, Myoglobin, and Cytochrom C as well as of the applied enzymes (glycosylated trypsin conjugates, methylated porcine trypsin and unmodified bovine trypsin) were determined by amino acid analysis performed in the laboratory of Dr. P. Hunziker at the University of Zürich. Isotopically enriched water (95% $H_2{}^{18}O$) used for preparation of internal peptide standards was from Sigma-Aldrich Chemie (Steinheim, Germany). Modified porcine trypsin was purchased from Promega (Mannheim, Germany), unmodified bovine trypsin from Roche Diagnostics (Basel, Switzerland). Dithiothreitol (DTT) and iodoacetamide (IAA) were obtained from Merck (Darmstadt, Germany). 1-Cyano-4-hydroxycinnamic acid (CHCA) was from Bruker Daltonik (Bremen, Germany).

4.1.2.2 Concept of the method

1. Gel bands containing 5 pmol of standard protein (BSA) were digested under tested conditions. Obtained peptides were extracted from the gel matrix and solvent was evaporated. For each quantification experiment at least three samples were prepared in parallel.

2. ^{18}O-labeled internal standards were obtained by digestion of the same standard protein (BSA) with known amount (0.3 pmol/µL) in a buffer containing $H_2{}^{18}O$ (95%), rendering tryptic peptides labeled with one or two ^{18}O atoms at their C-termini. Since yield of in-solution digestion is close to 100%, the protein concentration is directly proportional to the average concentration of individual tryptic peptides in the digest.

3. Tryptic peptides from in-gel digests were redissolved in 10 µL of internal standard and analyzed by MALDI TOF MS.

4. The yield of an individual peptide was calculated from the amount of the internal ^{18}O-labeled standard and the ratio of the area of the monoisotopic peak of the unlabeled peptide and of the sum of deconvoluted areas of monoisotopic peaks of singly and doubly ^{18}O-labeled forms of the internal standard.

4.1.2.3 Gel electrophoresis

One-dimensional SDS-polyacrylamide gel electrophoresis was performed as described [314] on the Bio-Rad Mini-Protean II system (Bio-Rad, Hercules, USA) using 10 and 12% polyacrylamide gels. Aliquots (5 pmol) of a standard protein (BSA, Aldolase, Myoglobin and Cytochrom C) were loaded onto each lane of the 7 x 10 cm minigel. Electrophoresis was conducted at a constant voltage of 150 V. After electrophoresis, protein bands were visualized by staining with Coomassie Brilliant Blue R-250 (Serva, Heidelberg, Germany). The bands were excised from the gel slab, cut into pieces, and put into 0.65-mL PCR microtubes.

4.1.2.4 In-gel digestion

Conventional digestion protocol (CDP) by unmodified trypsin. The digestion was carried out as described [137]. Proteins were in-gel reduced by 10mM dithiothreitol and alkylated by 55mM iodoacetamide. Destained, washed, dehydrated gel pieces were rehydrated for 60 minutes in ~0.5 µM solution of unmodified bovine trypsin in 25 mM ammonium bicarbonate buffer at 4°C and then digested overnight at 37°C.

Accelerated digestion protocol (ADP) by saccharide modified trypsin conjugates. To establish optimal conditions for accelerated in-gel digestion of proteins by glycosylated trypsin conjugates, the effect of enzyme concentration, digestion time and digestion temperature on the yield of digestion products was evaluated [2]. After the reduction/alkylation step dried gel pieces were rehydrated with trypsin conjugates at 4°C for 60 min; the enzyme concentration ranged from 0.5 to 3 µM in 25 mM ammonium bicarbonate. The digestion was performed at 55°C and 65°C for 30, 90 and 180 min.

Extraction of peptides. Peptides from the gel pieces were finally extracted with 5 % formic acid and acetonitrile as described [137] and the extracts were dried down in a vacuum centrifuge.

4.1.2.5 ^{18}O-labeled internal standards for quantification

9 µM BSA solution in 25 mM ammonium bicarbonate buffer in $H_2{}^{18}O$ (95%) was digested overnight at 37°C by unmodified bovine trypsin at an enzyme/substrate ratio

1:10 (w/w). The obtained stock solution of ^{18}O-labeled BSA peptides was diluted 30 times with $H_2{}^{18}O$ (95%) to get 0.3 µM internal standard. Tryptic peptides from in-gel digests were redissolved in 10 µL of internal standard and analyzed by MALDI TOF.

4.1.2.6 MALDI analysis

All experiments were performed using MALDI TOF instrument Reflex IV (Bruker Daltonik, Bremen, Germany), equipped with Scout 384 ion source. Spectra were processed by Xmass 5.1.1 and BioTools 2.1software (Bruker Daltonik). Proteins were identified using MASCOT software (version 2.1, Matrix Science, London, UK) installed on a local server; database searches were performed against a non-redundant protein database MSDB downloaded from the European Bioinformatics Institute (EBI).

A 1.2 µL aliquot of the sample was withdrawn onto the AnchorChip 600/384 target. 0.6 µL of the matrix solution (2 mg/ml 1-cyano-4-hydroxycinnamic acid in 2.5 % trifluoroacetic/acetonitrile, 1:2 v/v) was spiked directly into the analyte droplet. The mixture was allowed to dry down at room temperature, and the target was washed in 5% formic acid [292]. Each experiment was repeated 3 to 5 times; at least 3 samples were prepared for each experiment. Typically about 300 laser pulses per spectrum were accumulated and smoothed by Savitzky-Golay filter.

4.1.2.7 Deconvolution of isotopic clusters of ^{18}O-labeled peptides

Since the profile of the isotopic cluster contains singly and doubly O^{18}-labeled peptides a deconvolution method was applied, which uses the isotopic ratios calculated from the peptide composition. The amount of an individual (non-labeled) peptide (n_{16}) was calculated according equation 7:

$$n_{16} = \frac{{}^{1}A_1}{{}^{*3}A_1 + {}^{*2}A_1 \cdot (1 - f_3) + {}^{1}A_1 \cdot (f_3^2 - f_5 - f_3)} \cdot n_{18}$$

(Equation 7)

where n_{18} the amount of internal standard; ${}^{1}A_1$ is the monoisotopic peak area of the unlabeled peptide; ${}^{*3}A_1$, ${}^{*2}A_1$ are the peak areas of the first isotopic peaks of ^{18}O single labeled and double labeled peptides (spaced from the monoisotopic peak of the unlabeled peptide by 2 and 4 Da, respectively). The theoretic isotopic distributions for all peptides used in quantification experiments were calculated using PeptideProspector 4.0.4 (Univ. of California, http://prospector.ucsf.edu), presuming that the differences in isotopic distribution ratios for ^{18}O-labeled and unlabeled peptide are negligible. Coefficients f_3 and f_5 are the calculated ratios of the intensity of, respectively, third (+2 Da) and fifth (+4 Da) isotopic peaks to the intensity of the monoisotopic peak of the unlabeled peptide.

Equation 7 can be simplified to equation 8, when quantification is performed using arginine terminated peptides, which incorporate more than 90 % of ^{18}O in double-labeled form.

$$n_{16} = \frac{{}^{1}A_1}{{}^{*3}A_1} \cdot n_{18}$$

(Equation 8)

In the performed kinetic study the yield of the following three Arg-containing peptides was determined: YLYEIAR, m/z 927.49; LGEYGFQNALIVR, m/z 1479.79; and DAFLGSFLYEYSR m/z 1567.74. The yield of digestion was calculated by averaging the amount of these peptides measured in at least 3, parallel runs.

4.1.3 Study of digestion kinetics using label-free quantification approach

4.1.3.1 Chemicals

All chemicals were purchased from Sigma-Aldrich (Steinheim, Germany) and were of analytical grade, unless otherwise noted. Solvents for liquid chromatography were of Lichrosolv grade. Formic acid and trifluoroacetic acid (TFA) were purchased from Merck (Darmstadt, Germany). Modified porcine trypsin was purchased from Promega (Mannheim, Germany), unmodified bovine trypsin from Roche Diagnostics (Basel, Switzerland). Dithiothreitol (DTT) and iodoacetamide (IAA) were obtained from Merck (Darmstadt, Germany). Concentrations of stock solutions of the applied standard proteins (BSA, myosin, b-galactosidase, alcohol dehydrogenase, and myoglobin) as well as of the applied enzymes (glycosylated trypsin conjugates, methylated porcine trypsin and unmodified bovine trypsin) were determined by amino acid analysis performed in the laboratory of Dr. P. Hunziker at the University of Zürich.

4.1.3.2 Concept of the method

1. Gel bands containing 1 pmol BSA were digested under tested conditions, obtained peptides were extracted from the gel matrix and solvent was evaporated (as described in chapter 4.1.2). For each quantification experiment 3 samples were prepared in parallel.

2. Peptide mixture was redissolved in 10 µl of 0.05% TFA and 2 µl of the solution were subjected to nanoLC-MS/MS analysis. Each analysis was repeated three times to get better statistic of the measurements.

3. Proteolytic peptides were identified by correlating their fragmentation spectra with peptide sequences from MASCOT database.

4. For peptides subjected to quantification peak intensity areas were calculated by generating extracted ion chromatograms from the full scan mass spectra within a narrow m/z range, corresponding to different charge states of the peptides. To determine correct retention times of the quantified peptides scan numbers of the corresponding MS/MS spectra were used.

5. To quantify tryptic peptides produced by in-gel digestion of BSA in the kinetic experiments calibrating curves of the corresponding peptides were generated from in-solution digest of the same protein with known concentration. 7 aliquots of BSA in-solution digest containing protein amounts from 12 to 740 fmol (obtained in serial dilutions) were analyzed by nanoLC-MS/MS. Each analysis was repeated 3 times to ensure better statistic of the measurements. Regression lines were generated for each peptide subjected to quantification. The yield of digestion was calculated by averaging the amounts of the quantified peptides.

4.1.3.3 Redaction, alkylation and digestion of protein stock solutions

Stock solutions of the used standard proteins were prepared in 25 mM ammonium bicarbonate buffer at concentrations ranging from 50 to 200 µM. The obtained stock solutions were diluted down to the concentration of 10 µM and submitted to amino acid analysis in the laboratory of Dr. P. Hunziker at the University of Zürich, in order to determine their accurate protein concentration. Further each protein solution was reduced by adding dithiothreitol (DTT) to a final concentration of DTT ~ 1.5 mM and incubated for 30 min at 37°C. To alkulate the protein iodoacetamide (IAA) was added to the protein solution to a final concentration of ~ 3 mM and the mixture was incubated at the room temperature in the dark. The reduced and alkylated proteins were digested by addition of trypsin at the ratio of 1:50. The mixture was incubated for ca. 10 hours at 37°C. Several aliquots containing 10 µL of each digest were withdrawn and put into HPLC vials; the solvent was evaporated in a vacuum centrifuge and the peptide mixture was stored at -20°C.

4.1.3.4 NanoLC- MS/MS analysis

Aliquots of sample digests dissolved in 0.05% TFA were injected into a nanoLC-MS/MS Ultimate system (Dionex, Amsterdam, The Netherlands) interfaced on-line to a linear ion trap LTQ (ThermoElectron Corp., San Jose, CA). Peptides were first loaded onto a 1 mm × 300 µm i.d. trapping microcolumn packed with C18 PepMAP100 5µm particles (Dionex) in 0.05% TFA at the flow rate of 20µL/min. After a 4 min wash, they were back-flush-eluted and separated on a 15 cm × 75 µm i.d. nanocolumn packed with

C18 PepMAP100 3 µm particles (Dionex) at the flow rate of 200 nL/min using the following mobile phase gradient: from 5 to 20% of solvent B in 20 min, 20-50% B in 16 min, 50-100% B in 5 min, 100% B during 10 min, and back to 5% B in 5 min. Solvent A was 95:5 H_2O/acetonitrile (v/v) with 0.1% formic acid; solvent B was 20:80 H_2O/acetonitrile (v/v) with 0.1% formic acid. Peptides were eluted into the mass spectrometer via a dynamic nanospray probe (Thermo Electron Corp.). A silicatip uncoated needle (20 µm i.d., 10 µm tip ID) (New Objective, Woburn, MA) was used with a spray voltage of 1.8 kV, and the transfer capillary temperature was set at 200°C. Data-dependent acquisition was controlled by Xcalibur 1.4 software (ThermoElectron Corp.). The acquisition cycle consisted of a survey scan covering the range of m/z 350-1500 followed by MS/MS fragmentation of the three most intense precursor ions under the relative collision energy of 35%, triggered by a minimum signal threshold of 500 counts with the isolation width of 4.0 amu. Spectra were acquired under automated gain control (AGC) in three microscans for survey spectra and for MS/MS spectra, with maximal ion injection time of 100 ms. The m/z of fragmented precursor ions were dynamically excluded for a further 60 s, but otherwise no pre-defined exclusion lists were applied. Spectra were exported as dta files using BioWorks 3.1 software (Thermo Electron Corp.) under the following settings: peptide mass range, 500-3500; minimum total ion intensity threshold, 1000; minimum number of fragment ions, 15; precursor mass tolerance, 1.4 amu; group scan, 1; minimum group count, 1.

4.1.3.5 MASCOT database searches

Tandem mass spectra were searched against an MSDB database (updated May 15, 2005; contains 2 011 425 protein sequence entries) by MASCOT v. 2.1 software (Matrix Science, London, UK) installed on a local 2 CPU server. Mass tolerance for precursor and fragment ions was 2.0 and 0.5 Da, respectively. Other search parameters were: instrument profile, ESI-Trap; fixed modification, carbamidomethyl (cysteine); variable modification, oxidation (methionine).

4.1.3.6 Peak extraction

Peptide ion intensities were calculated using extracted ion chromatograms (XICs), which were generated from the full scan mass spectra within a narrow m/z range, corresponding to different charge states of the analyzed peptide (triple, double, and single). Differently charged ions of each analyzed peptide were extracted from the full scan mass spectra using a lower limit of ***m/z = ((peptide monoisotopic mass – 1) + charge)/charge*** and upper limit of ***m/z = ((peptide monoisotopic mass + 2) + charge)/charge***. The ion intensity of a peptide was subsequently calculated by summing peak areas of its triple, double, and single charged ions. XICs were generated using Xcalibur1.4 software (ThermoElectron Corp., San Jose, CA), which offers peak finding, peak smoothing, and integration functions. To enable correct peak finding scan numbers of the corresponding MS/MS spectra (confidently identified by MASCOT search) were used.

4.1.3.7 Quantification of in-gel digestion products

To quantify tryptic peptides obtained by in-gel digestion of BSA in the performed kinetic study calibrating curves of the corresponding peptides from in-solution digest of the same standard protein were generated. To this end a 740 µM stock solution of BSA was digested as described in chapter 4.1.3.3. 10 µL of this stock solution were withdrawn and put into HPLC vial, the solvent was evaporated in the vacuum centrifuge and the peptide mixture was redissolved in 100 µl of 0.05% TFA. The obtained solution was used to prepare a dilution series including 7 mixtures. 2 µl aliquots of these mixtures containing 12, 23, 46, 93, 185, 370, and 740 fmol of BSA were subjected to nanoLC-MS/MS analysis. Each measurement was repeated three times to ensure better statistic of the acquisition. Subsequently regression lines for each subjected to quantification peptide were generated.

The concentrations of six tryptic peptides, AEFVEVTK (M = 921.48), YLYEIAR (M = 926.49), KQTALVELLK (M = 1141.71), LVNELTEFAK (M = 1162.62), HLVDEPQNLIK (M = 1304.71), and KVPQVSTPTLVEVSR (M = 1638.93), were determined for each kinetic experiment, and the results were averaged.

4.2 Validation of protein identifications with borderline statistical confidence

4.2.1 Chemicals

All chemicals were purchased from Sigma-Aldrich (Steinheim, Germany) and were of analytical grade, unless otherwise noted. Solvents for liquid chromatography were of Lichrosolv grade; formic and trifluoroacetic acids were purchased from Merck (Darmstadt, Germany).

4.2.2 Protein datasets

Proteins were isolated from *Caenorhabditis elegans* worms in two collaboration projects with Prof. A. Hyman's laboratory (MPI-CBG, Dresden) and purified by affinity chromatography as described [4], [5]. Purified proteins were then separated by one-dimensional SDS-polyacrylamide gel electrophoresis; protein bands were visualized by Coomassie Brilliant Blue R250 staining and excised from the gel matrix. The excised bands were in-gel-digested with trypsin[137]. Tryptic peptides, recovered from the gel pieces by extraction with 5% formic acid and acetonitrile, were dried in a vacuum centrifuge and stored at -20°C until analyzed.

4.2.3 NanoLC-MS/MS analysis

Nano LC-MS/MS analysis was carried out as previously described in chapter 4.1.3.4 only with a difference that MS/MS fragmentation was performed on five most intense precursor ions and spectra were acquired (under AGC) in one microscan for survey spectra and three microscans for MS/MS spectra.

4.2.4 MASCOT database searches

Tandem mass spectra were searched against an MSDB database (updated May 15, 2005; contains 2 011 425 protein sequence entries) by MASCOT v. 2.1 software (Matrix Science, London, UK) installed on a local 2 CPU server. Mass tolerance for precursor and fragment ions was 2.0 and 0.5 Da, respectively. Other search parameter were: instrument profile, ESI-Trap; fixed modification, carbamidomethyl (cysteine); variable modification, oxidation (methionine). Where specified, searches were performed against a subset of *C. elegans* proteins that comprised 30 304 protein sequence entries. Hits were regarded as confident if more than three peptides were matched by MASCOT search with ions scores above the confidence threshold for a species-specific database (36 for *C. elegans* protein database), or at least one score was above the threshold for a comprehensive database (53 for MSDB). Hits were regarded as borderline if MASCOT matched less than four peptides and the ion score of at least one peptide was within the range of ±30% of the threshold score (from 26 to 46 for *C. elegans*).

4.2.5 *De Novo* peptide sequencing and MS BLAST searches

Where specified, files in dta format were converted into MASCOT generic format (mgf) and sequenced *de novo* by a modified version of PepNovo software [179] installed on a desktop (Pentium IV) PC. A single MS/MS spectrum was typically interpreted *de novo* in less than 0.5 s, and up to seven partially redundant candidate sequences were produced. To each interpreted spectrum, PepNovo assigned a quality score, which stands for the expected number of confidently determined amino acid residues in the most accurate sequence proposal. This score was derived from the sum of the probabilities of the individual amino acids being correct, which were computed using a logistic regression model [315]. Candidate sequences were then edited according to MS BLAST conventions and merged into a single search string in arbitrary order [220, 224, 225]. MS BLAST searches were performed against nr database at http://genetics.bwh.harvard.edu/msblast/ under the following settings: Scoring Table, 99; Filter, none; Expect, 1000. Statistical significance of hits was evaluated according to MS BLAST scoring scheme [225]. A typical search with a query of seven candidate sequences required less than 15 s to complete.

4.2.6 PepNovo/MS BLAST validation performance

The entire procedure was performed using script written by Henrik Thomas. A simulation dataset was built out of 100 high-quality peptide spectra, each represented by a single dta file. Upon MASCOT database search, each spectrum unequivocally hit a single peptide sequence of, on average, 12 amino acid residues with the ions scores above 70. In each spectrum, peaks with relative intensities below 1% of the base peak intensity were declared noise, and their absolute intensity was left unchanged, whereas a dedicated script reduced the absolute intensity of other peaks with the step of 1% and, hence, produced the series of 100 spectra with the gradually altered signal-to-noise ratios. Their dta files were merged into a single mgf file and submitted to MASCOT search, and ions scores of spectra matched to the correct database peptide sequence were registered. In parallel, the same mgf file was sequenced *de novo* by the PepNovo program in a batch mode, recording up to seven sequence candidates for each interpreted spectrum. PepNovo scores of predicted sequences were registered, and sequences were merged into query strings and submitted to MS BLAST searches. The outcome of MS BLAST searches was sorted into three groups as follows: where MS BLAST produced a hit that was also confident according to MS BLAST scoring scheme (first group); where the target peptide was listed in the output of the MS BLAST search as a borderline or nonconfident hit (second group); or where the target protein was not hit by MS BLAST at all (third group). In each series, I aimed to identify (if possible) the two spectra with the lowest signal-to-noise ratios that belonged to the first and second groups and registered their ions scores (MASCOT), sequence quality scores (PepNovo), and MS BLAST scores (solely for the reference). The same simulation routine was applied to all 100 high-quality spectra from the initial dataset.

REFERENCES

1. Sebela M, Stosova T, Havlis J, Wielsch N, Thomas H, Zdrahal Z, Shevchenko A: **Thermostable trypsin conjugates for high-throughput proteomics: synthesis and performance evaluation**. *Proteomics* 2006, **6**(10):2959-2963.

2. Havlis J, Thomas H, Sebela M, Shevchenko A: **Fast-response proteomics by accelerated in-gel digestion of proteins**. *Anal Chem* 2003, **75**(6):1300-1306.

3. Wielsch N, Thomas H, Surendranath V, Waridel P, Frank A, Pevzner P, Shevchenko A: **Rapid validation of protein identifications with the borderline statistical confidence via de novo sequencing and MS BLAST searches**. *J Proteome Res* 2006, **5**(9):2448-2456.

4. Ozlu N, Srayko M, Kinoshita K, Habermann B, O'Toole E T, Muller-Reichert T, Schmalz N, Desai A, Hyman AA: **An essential function of the C. elegans ortholog of TPX2 is to localize activated aurora A kinase to mitotic spindles**. *Dev Cell* 2005, **9**(2):237-248.

5. Schlaitz AL, Srayko M, Dammermann A, Quintin S, Wielsch N, MacLeod I, de Robillard Q, Zinke A, Yates JR, 3rd, Muller-Reichert T *et al*: **The C. elegans RSA complex localizes protein phosphatase 2A to centrosomes and regulates mitotic spindle assembly**. *Cell* 2007, **128**(1):115-127.

6. Vidal M: **A biological atlas of functional maps**. *Cell* 2001, **104**(3):333-339.

7. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM: **A biochemical genomics approach for identifying genes by the activity of their products**. *Science* 1999, **286**(5442):1153-1155.

8. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al*: **Global analysis of protein activities using proteome chips**. *Science* 2001, **293**(5537):2101-2105.

9. MacBeath G, Schreiber SL: **Printing proteins as microarrays for high-throughput function determination**. *Science* 2000, **289**(5485):1760-1763.

10. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al*: **The genome sequence of Drosophila melanogaster**. *Science* 2000, **287**(5461):2185-2195.

11. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**(6915):520-562.

12. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A *et al*: **The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics**. *PLoS Biol* 2003, **1**(2):E45.

13. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.

14. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

15. **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.

16. Aebersold R: **Constellations in a cellular universe**. *Nature* 2003, **422**(6928):115-116.

17. Patterson SD, Aebersold RH: **Proteomics: the first decade and beyond**. *Nat Genet* 2003, **33 Suppl**:311-323.

18. Aebersold R, Goodlett DR: **Mass spectrometry in proteomics**. *Chem Rev* 2001, **101**(2):269-295.

19. Pandey A, Mann M: **Proteomics to study genes and genomes**. *Nature* 2000, **405**(6788):837-846.

20. Phizicky E, Bastiaens PI, Zhu H, Snyder M, Fields S: **Protein analysis on a proteomic scale**. *Nature* 2003, **422**(6928):208-215.

21. Domon B, Aebersold R: **Mass spectrometry and protein analysis**. *Science* 2006, **312**(5771):212-217.

22. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM: **Electrospray ionization for mass spectrometry of large biomolecules**. *Science* 1989, **246**(4926):64-71.

23. Karas M, Hillenkamp F: **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons**. *Anal Chem* 1988, **60**(20):2299-2301.

24. Cech NB, Enke CG: **Relating electrospray ionization response to nonpolar character of small peptides**. *Anal Chem* 2000, **72**(13):2717-2723.

25. Krause E, Wenschuh H, Jungblut PR: **The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins**. *Anal Chem* 1999, **71**(19):4160-4165.

26. Stapels MD, Barofsky DF: **Complementary use of MALDI and ESI for the HPLC-MS/MS analysis of DNA-binding proteins**. *Anal Chem* 2004, **76**(18):5423-5430.

27. Getie M, Schmelzer CE, Weiss AS, Neubert RH: **Complementary mass spectrometric techniques to achieve complete sequence coverage of recombinant human tropoelastin**. *Rapid Commun Mass Spectrom* 2005, **19**(20):2989-2993.

28. Smith SA, Blake TA, Ifa DR, Cooks RG, Ouyang Z: **Dual-source mass spectrometer with MALDI-LIT-ESI configuration**. *J Proteome Res* 2007, **6**(2):837-845.

29. Bodnar WM, Blackburn RK, Krise JM, Moseley MA: **Exploiting the complementary nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for increased proteome coverage**. *J Am Soc Mass Spectrom* 2003, **14**(9):971-979.

30. Wysocki VH, Resing KA, Zhang Q, Cheng G: **Mass spectrometry of peptides and proteins**. *Methods* 2005, **35**(3):211-222.

31. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**(6928):198-207.

32. Yost RA, Boyd RK: **Tandem mass spectrometry: quadrupole and hybrid instruments**. *Methods Enzymol* 1990, **193**:154-200.

33. Cotter RJ: **Time-of-flight mass spectrometry for the structural analysis of biological molecules**. *Anal Chem* 1992, **64**(21):1027A-1039A.

34. Chernushevich IV, Loboda AV, Thomson BA: **An introduction to quadrupole-time-of-flight mass spectrometry**. *J Mass Spectrom* 2001, **36**(8):849-865.

35. Roepstorff P: **MALDI-TOF mass spectrometry in protein chemistry**. *Exs* 2000, **88**:81-97.

36. Pappin DJ: **Peptide mass fingerprinting using MALDI-TOF mass spectrometry**. *Methods Mol Biol* 1997, **64**:165-173.

37. Pappin DJ: **Peptide mass fingerprinting using MALDI-TOF mass spectrometry**. *Methods Mol Biol* 2003, **211**:211-219.

38. Spengler B, Kirsch D, Kaufmann R, Jaeger E: **Peptide sequencing by matrix-assisted laser-desorption mass spectrometry**. *Rapid Commun Mass Spectrom* 1992, **6**(2):105-108.

39. Kaufmann R, Chaurand P, Kirsch D, Spengler B: **Post-source decay and delayed extraction in matrix-assisted laser desorption/ionization-reflectron time-of-flight mass spectrometry. Are there trade-offs?** *Rapid Commun Mass Spectrom* 1996, **10**(10):1199-1208.

40. Schnaible V, Wefing S, Resemann A, Suckau D, Bucker A, Wolf-Kummeth S, Hoffmann D: **Screening for disulfide bonds in proteins by MALDI in-source decay and LIFT-TOF/TOF-MS**. *Anal Chem* 2002, **74**(19):4980-4988.

41. Kenny DJ, Brown JM, Palmer ME, Snel MF, Bateman RH: **A parallel approach to post source decay MALDI-TOF analysis**. *J Am Soc Mass Spectrom* 2006, **17**(1):60-66.

42. Medzihradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML, Burlingame AL: **The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer**. *Anal Chem* 2000, **72**(3):552-558.

43. Baldwin MA: **Mass spectrometers for the analysis of biomolecules**. *Methods Enzymol* 2005, **402**:3-48.

44. Ens W, Standing KG: **Hybrid quadrupole/time-of-flight mass spectrometers for analysis of biomolecules**. *Methods Enzymol* 2005, **402**:49-78.

45. Shevchenko A, Loboda A, Ens W, Standing KG: **MALDI quadrupole time-of-flight mass spectrometry: a powerful tool for proteomic research**. *Anal Chem* 2000, **72**(9):2132-2141.

46. Krutchinsky AN, Zhang W, Chait BT: **Rapidly switchable matrix-assisted laser desorption/ionization and electrospray quadrupole-time-of-flight mass**

spectrometry for protein identification. *J Am Soc Mass Spectrom* 2000, **11**(6):493-504.

47. Cha B, Blades M, Douglas DJ: **An interface with a linear quadrupole ion guide for an electrospray-ion trap mass spectrometer system**. *Anal Chem* 2000, **72**(22):5647-5654.

48. Jonscher KR, Yates JR, 3rd: **The quadrupole ion trap mass spectrometer--a small solution to a big challenge**. *Anal Biochem* 1997, **244**(1):1-15.

49. Schwartz JC, Senko MW, Syka JE: **A two-dimensional quadrupole ion trap mass spectrometer**. *J Am Soc Mass Spectrom* 2002, **13**(6):659-669.

50. Mayya V, Rezaul K, Cong YS, Han D: **Systematic comparison of a two-dimensional ion trap and a three-dimensional ion trap mass spectrometer in proteomics**. *Mol Cell Proteomics* 2005, **4**(2):214-223.

51. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M: **Global, in vivo, and site-specific phosphorylation dynamics in signaling networks**. *Cell* 2006, **127**(3):635-648.

52. Douglas DJ, Frank AJ, Mao D: **Linear ion traps in mass spectrometry**. *Mass Spectrom Rev* 2005, **24**(1):1-29.

53. Hager JW, Yves Le Blanc JC: **Product ion scanning using a Q-q-Q linear ion trap (Q TRAP) mass spectrometer**. *Rapid Commun Mass Spectrom* 2003, **17**(10):1056-1064.

54. Le Blanc JC, Hager JW, Ilisiu AM, Hunter C, Zhong F, Chu I: **Unique scanning capabilities of a new hybrid linear ion trap mass spectrometer (Q TRAP) used for high sensitivity proteomics applications**. *Proteomics* 2003, **3**(6):859-869.

55. Marshall AG, Hendrickson CL, Jackson GS: **Fourier transform ion cyclotron resonance mass spectrometry: a primer**. *Mass Spectrom Rev* 1998, **17**(1):1-35.

56. Comisarow MB, Marshall AG: **The early development of Fourier transform ion cyclotron resonance (FT-ICR) spectroscopy**. *J Mass Spectrom* 1996, **31**(6):581-585.

57. Emmett MR, White FM, Hendrickson CL, Shi SD, Marshall AG: **Application of micro-electrospray liquid chromatography techniques to FT-ICR MS to enable high-sensitivity biological analysis**. *J Am Soc Mass Spectrom* 1998, **9**(4):333-340.

58. Wilcox BE, Hendrickson CL, Marshall AG: **Improved ion extraction from a linear octopole ion trap: SIMION analysis and experimental demonstration**. *J Am Soc Mass Spectrom* 2002, **13**(11):1304-1312.

59. Bogdanov B, Smith RD: **Proteomics by FTICR mass spectrometry: top down and bottom up**. *Mass Spectrom Rev* 2005, **24**(2):168-200.

60. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R: **The Orbitrap: a new mass spectrometer**. *J Mass Spectrom* 2005, **40**(4):430-443.

61. Hardman M, Makarov AA: **Interfacing the orbitrap mass analyzer to an electrospray ion source**. *Anal Chem* 2003, **75**(7):1699-1705.

62. Makarov A: **Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis**. *Anal Chem* 2000, **72**(6):1156-1162.

63. Makarov A, Denisov E, Kholomeev A, Balschun W, Lange O, Strupat K, Horning S: **Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer**. *Anal Chem* 2006, **78**(7):2113-2120.

64. Scigelova M, Makarov A: **Orbitrap mass analyzer - overview and applications in proteomics**. *Proteomics* 2006, **6 Suppl 2**:16-21.

65. Makarov A, Denisov E, Lange O, Horning S: **Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer**. *J Am Soc Mass Spectrom* 2006, **17**(7):977-982.

66. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, Makarov A, Lange O, Horning S, Mann M: **Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap**. *Mol Cell Proteomics* 2005, **4**(12):2010-2021.

67. Yates JR, Cociorva D, Liao L, Zabrouskov V: **Performance of a linear ion trap-Orbitrap hybrid for peptide analysis**. *Anal Chem* 2006, **78**(2):493-500.

68. Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M: **Higher-energy C-trap dissociation for peptide modification analysis**. *Nat Methods* 2007, **4**(9):709-712.

69. Macek B, Waanders LF, Olsen JV, Mann M: **Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer**. *Mol Cell Proteomics* 2006, **5**(5):949-958.

70. Anderson NL, Anderson NG: **The human plasma proteome: history, character, and diagnostic prospects**. *Mol Cell Proteomics* 2002, **1**(11):845-867.

71. Fournier ML, Gilmore JM, Martin-Brown SA, Washburn MP: **Multidimensional separations-based shotgun proteomics**. *Chem Rev* 2007, **107**(8):3654-3686.

72. Premstaller A, Oberacher H, Walcher W, Timperio AM, Zolla L, Chervet JP, Cavusoglu N, van Dorsselaer A, Huber CG: **High-performance liquid chromatography-electrospray ionization mass spectrometry using monolithic capillary columns for proteomic studies**. *Anal Chem* 2001, **73**(11):2390-2396.

73. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR, 3rd: **Direct analysis of protein complexes using mass spectrometry**. *Nat Biotechnol* 1999, **17**(7):676-682.

74. Xie H, Bandhakavi S, Griffin TJ: **Evaluating preparative isoelectric focusing of complex peptide mixtures for tandem mass spectrometry-based proteomics: a case study in profiling chromatin-enriched subcellular fractions in Saccharomyces cerevisiae**. *Anal Chem* 2005, **77**(10):3198-3207.

75. Malmstrom J, Lee H, Nesvizhskii AI, Shteynberg D, Mohanty S, Brunner E, Ye M, Weber G, Eckerskorn C, Aebersold R: **Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis**. *J Proteome Res* 2006, **5**(9):2241-2249.

76. Heller M, Michel PE, Morier P, Crettaz D, Wenz C, Tissot JD, Reymond F, Rossier JS: **Two-stage Off-Gel isoelectric focusing: protein followed by peptide fractionation and application to proteome analysis of human plasma**. *Electrophoresis* 2005, **26**(6):1174-1188.

77. Heller M, Ye M, Michel PE, Morier P, Stalder D, Junger MA, Aebersold R, Reymond F, Rossier JS: **Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides**. *J Proteome Res* 2005, **4**(6):2273-2282.

78. Shen Y, Smith RD, Unger KK, Kumar D, Lubda D: **Ultrahigh-throughput proteomics using fast RPLC separations with ESI-MS/MS**. *Anal Chem* 2005, **77**(20):6692-6701.

79. Yin H, Killeen K, Brennen R, Sobek D, Werlich M, van de Goor T: **Microfluidic chip for peptide analysis with an integrated HPLC column, sample enrichment column, and nanoelectrospray tip**. *Anal Chem* 2005, **77**(2):527-533.

80. Simpson DC, Ahn S, Pasa-Tolic L, Bogdanov B, Mottaz HM, Vilkov AN, Anderson GA, Lipton MS, Smith RD: **Using size exclusion chromatography-RPLC and RPLC-CIEF as two-dimensional separation strategies for protein profiling**. *Electrophoresis* 2006, **27**(13):2722-2733.

81. Mohan D, Pasa-Tolic L, Masselon CD, Tolic N, Bogdanov B, Hixson KK, Smith RD, Lee CS: **Integration of electrokinetic-based multidimensional separation/concentration platform with electrospray ionization-Fourier transform ion cyclotron resonance-mass spectrometry for proteome analysis of Shewanella oneidensis**. *Anal Chem* 2003, **75**(17):4432-4440.

82. Peng J, Gygi SP: **Proteomics: the move to mixtures**. *J Mass Spectrom* 2001, **36**(10):1083-1091.

83. Malmstrom J, Lee H, Aebersold R: **Advances in proteomic workflows for systems biology**. *Curr Opin Biotechnol* 2007, **18**(4):378-384.

84. Resing KA, Ahn NG: **Proteomics strategies for protein identification**. *FEBS Lett* 2005, **579**(4):885-889.

85. Chait BT: **Chemistry. Mass spectrometry: bottom-up or top-down?** *Science* 2006, **314**(5796):65-66.

86. Han X, Jin M, Breuker K, McLafferty FW: **Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons**. *Science* 2006, **314**(5796):109-112.

87. Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, Lewis MA, Carpenter BK, McLafferty FW: **Electron capture dissociation for structural characterization of multiply charged protein cations**. *Anal Chem* 2000, **72**(3):563-573.

88. Coon JJ, Ueberheide B, Syka JE, Dryhurst DD, Ausio J, Shabanowitz J, Hunt DF: **Protein identification using sequential ion/ion reactions and tandem mass spectrometry**. *Proc Natl Acad Sci U S A* 2005, **102**(27):9463-9468.

89. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF: **Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry**. *Proc Natl Acad Sci U S A* 2004, **101**(26):9528-9533.

90. Forbes AJ, Patrie SM, Taylor GK, Kim YB, Jiang L, Kelleher NL: **Targeted analysis and discovery of posttranslational modifications in proteins from methanogenic archaea by top-down MS**. *Proc Natl Acad Sci U S A* 2004, **101**(9):2678-2683.

91. Kelleher NL: **Top-down proteomics**. *Anal Chem* 2004, **76**(11):197A-203A.

92. Du Y, Parks BA, Sohn S, Kwast KE, Kelleher NL: **Top-down approaches for measuring expression ratios of intact yeast proteins using Fourier transform mass spectrometry**. *Anal Chem* 2006, **78**(3):686-694.

93. Parks BA, Jiang L, Thomas PM, Wenger CD, Roth MJ, Ii MT, Burke PV, Kwast KE, Kelleher NL: **Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers**. *Anal Chem* 2007, **79**(21):7984-7991.

94. Li W, Hendrickson CL, Emmett MR, Marshall AG: **Identification of intact proteins in mixtures by alternated capillary liquid chromatography electrospray ionization and LC ESI infrared multiphoton dissociation Fourier transform ion cyclotron resonance mass spectrometry**. *Anal Chem* 1999, **71**(19):4397-4402.

95. Waanders LF, Hanke S, Mann M: **Top-Down Quantitation and Characterization of SILAC-Labeled Proteins**. *J Am Soc Mass Spectrom* 2007, **18**(11):2058-2064.

96. Reid GE, McLuckey SA: **'Top down' protein characterization via tandem mass spectrometry**. *J Mass Spectrom* 2002, **37**(7):663-675.

97. Taylor GK, Kim YB, Forbes AJ, Meng F, McCarthy R, Kelleher NL: **Web and database software for identification of intact proteins using "top down" mass spectrometry**. *Anal Chem* 2003, **75**(16):4081-4086.

98. LeDuc RD, Taylor GK, Kim YB, Januszyk TE, Bynum LH, Sola JV, Garavelli JS, Kelleher NL: **ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W340-345.

99. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL: **ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W701-706.

100. Horn DM, Zubarev RA, McLafferty FW: **Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules**. *J Am Soc Mass Spectrom* 2000, **11**(4):320-332.

101. Henzel WJ, Watanabe C, Stults JT: **Protein identification: the origins of peptide mass fingerprinting**. *J Am Soc Mass Spectrom* 2003, **14**(9):931-942.

102. Sommerer N, Centeno D, Rossignol M: **Peptide mass fingerprinting: identification of proteins by MALDI-TOF**. *Methods Mol Biol* 2007, **355**:219-234.

103. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R: **Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology**. *Proc Natl Acad Sci U S A* 2000, **97**(17):9390-9395.

104. Gygi SP, Aebersold R: **Mass spectrometry and proteomics**. *Curr Opin Chem Biol* 2000, **4**(5):489-494.

105. Wolters DA, Washburn MP, Yates JR, 3rd: **An automated multidimensional protein identification technology for shotgun proteomics**. *Anal Chem* 2001, **73**(23):5683-5690.

106. Washburn MP, Wolters D, Yates JR, 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology**. *Nat Biotechnol* 2001, **19**(3):242-247.

107. Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M: **In-gel digestion for mass spectrometric characterization of proteins and proteomes**. *Nat Protoc* 2006, **1**(6):2856-2860.

108. Giddings JC: **Two-dimensional separations: concept and promise**. *Anal Chem* 1984, **56**(12):1258A-1260A, 1262A, 1264A passim.

109. Davis JM, Giddings JC: **Origin and characterization of departures from the statistical model of component-peak overlap in chromatography**. *J Chromatogr* 1984, **289**:277-298.

110. Davis JM, Giddings JC: **Statistical method for estimation of number of components from single complex chromatograms: application to experimental chromatograms**. *Anal Chem* 1985, **57**(12):2178-2182.

111. McDonald WH, Yates JR, 3rd: **Shotgun proteomics and biomarker discovery**. *Dis Markers* 2002, **18**(2):99-105.

112. Florens L, Washburn MP: **Proteomic analysis by multidimensional protein identification technology**. *Methods Mol Biol* 2006, **328**:159-175.

113. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL *et al*: **A proteomic view of the Plasmodium falciparum life cycle**. *Nature* 2002, **419**(6906):520-526.

114. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL *et al*: **Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii**. *Nature* 2002, **419**(6906):512-519.

115. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK *et al*: **A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses**. *Science* 2005, **307**(5706):82-86.

116. Tomomori-Sato C, Sato S, Parmely TJ, Banks CA, Sorokina I, Florens L, Zybailov B, Washburn MP, Brower CS, Conaway RC *et al*: **A mammalian mediator subunit that shares properties with Saccharomyces cerevisiae mediator subunit Cse2**. *J Biol Chem* 2004, **279**(7):5846-5851.

117. Sato S, Tomomori-Sato C, Parmely TJ, Florens L, Zybailov B, Swanson SK, Banks CA, Jin J, Cai Y, Washburn MP *et al*: **A set of consensus mammalian mediator subunits identified by multidimensional protein identification technology**. *Mol Cell* 2004, **14**(5):685-691.

118. Powell DW, Weaver CM, Jennings JL, McAfee KJ, He Y, Weil PA, Link AJ: **Cluster analysis of mass spectrometry data reveals a novel component of SAGA**. *Mol Cell Biol* 2004, **24**(16):7249-7259.

119. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP: **Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae**. *J Proteome Res* 2006, **5**(9):2339-2347.

120. Washburn MP, Ulaszek RR, Yates JR, 3rd: **Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology**. *Anal Chem* 2003, **75**(19):5054-5061.

121. Zybailov B, Coleman MK, Florens L, Washburn MP: **Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling**. *Anal Chem* 2005, **77**(19):6218-6224.

122. MacCoss MJ, Wu CC, Liu H, Sadygov R, Yates JR, 3rd: **A correlation algorithm for the automated quantitative analysis of shotgun proteomics data**. *Anal Chem* 2003, **75**(24):6912-6921.

123. Shen Y, Tolic N, Masselon C, Pasa-Tolic L, Camp DG, 2nd, Lipton MS, Anderson GA, Smith RD: **Nanoscale proteomics**. *Anal Bioanal Chem* 2004, **378**(4):1037-1045.

124. Hunt DF, Michel H, Dickinson TA, Shabanowitz J, Cox AL, Sakaguchi K, Appella E, Grey HM, Sette A: **Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad**. *Science* 1992, **256**(5065):1817-1820.

125. MacNair JE, Patel KD, Jorgenson JW: **Ultrahigh-pressure reversed-phase capillary liquid chromatography: isocratic and gradient elution using columns packed with 1.0-micron particles**. *Anal Chem* 1999, **71**(3):700-708.

126. MacNair JE, Lewis KC, Jorgenson JW: **Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns**. *Anal Chem* 1997, **69**(6):983-989.

127. Xiang Y, Yan B, Yue B, McNeff CV, Carr PW, Lee ML: **Elevated-temperature ultrahigh-pressure liquid chromatography using very small polybutadiene-coated nonporous zirconia particles**. *J Chromatogr A* 2003, **983**(1-2):83-89.

128. Mellors JS, Jorgenson JW: **Use of 1.5-microm porous ethyl-bridged hybrid particles as a stationary-phase support for reversed-phase ultrahigh-pressure liquid chromatography**. *Anal Chem* 2004, **76**(18):5441-5450.

129. Xiang Y, Yan B, McNeff CV, Carr PW, Lee ML: **Synthesis of micron diameter polybutadiene-encapsulated non-porous zirconia particles for ultrahigh pressure liquid chromatography**. *J Chromatogr A* 2003, **1002**(1-2):71-78.

130. Motoyama A, Venable JD, Ruse CI, Yates JR, 3rd: **Automated ultra-high-pressure multidimensional protein identification technology (UHP-MudPIT) for improved peptide identification of proteomic samples**. *Anal Chem* 2006, **78**(14):5109-5118.

131. Yang Y, Thannhauser TW, Li L, Zhang S: **Development of an integrated approach for evaluation of 2-D gel image analysis: impact of multiple proteins in single spots on comparative proteomics in conventional 2-D gel/MALDI workflow**. *Electrophoresis* 2007, **28**(12):2080-2094.

132. Nicholas B, Skipp P, Mould R, Rennard S, Davies DE, O'Connor CD, Djukanovic R: **Shotgun proteomic analysis of human-induced sputum**. *Proteomics* 2006, **6**(15):4390-4401.

133. Rezaul K, Wu L, Mayya V, Hwang SI, Han D: **A systematic characterization of mitochondrial proteome from human T leukemia cells**. *Mol Cell Proteomics* 2005, **4**(2):169-181.

134. Zhu W, Venable J, Giometti CS, Khare T, Tollaksen S, Ahrendt AJ, Yates JR, 3rd: **Large-scale muLC-MS/MS for silver- and Coomassie blue-stained polyacrylamide gels**. *Electrophoresis* 2005, **26**(23):4495-4507.

135. Delahunty C, Yates JR, 3rd: **Protein identification using 2D-LC-MS/MS**. *Methods* 2005, **35**(3):248-255.

136. Shevchenko A, Loboda A, Ens W, Schraven B, Standing KG: **Archived polyacrylamide gels as a resource for proteome characterization by mass spectrometry**. *Electrophoresis* 2001, **22**(6):1194-1203.

137. Shevchenko A, Wilm M, Vorm O, Mann M: **Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels**. *Anal Chem* 1996, **68**(5):850-858.

138. Venkatesh R, Sundaram PV: **Modulation of stability properties of bovine trypsin after in vitro structural changes with a variety of chemical modifiers**. *Protein Eng* 1998, **11**(8):691-698.

139. Murphy A, C OF: **Chemically stabilized trypsin used in dipeptide synthesis**. *Biotechnol Bioeng* 1998, **58**(4):366-373.

140. Villalonga R, Fernandez M, Fragoso A, Cao R, Mariniello L, Porta R: **Thermal stabilization of trypsin by enzymic modification with beta-cyclodextrin derivatives**. *Biotechnol Appl Biochem* 2003, **38**(Pt 1):53-59.

141. Villalonga ML, Reyes G, Fragoso A, Cao R, Fernandez L, Villalonga R: **Chemical glycosidation of trypsin with O-carboxymethyl-poly-beta-cyclodextrin: catalytic and stability properties**. *Biotechnol Appl Biochem* 2005, **41**(Pt 3):217-223.

142. Bark SJ, Muster N, Yates JR, 3rd, Siuzdak G: **High-temperature protein mass mapping using a thermophilic protease**. *J Am Chem Soc* 2001, **123**(8):1774-1775.

143. Wang C, Eufemi M, Turano C, Giartosio A: **Influence of the carbohydrate moiety on the stability of glycoproteins**. *Biochemistry* 1996, **35**(23):7299-7307.

144. Mozhaev VV, Siksnis VA, Melik-Nubarov NS, Galkantaite NZ, Denis GJ, Butkus EP, Zaslavsky B, Mestechkina NM, Martinek K: **Protein stabilization via**

hydrophilization. Covalent modification of trypsin and alpha-chymotrypsin. *Eur J Biochem* 1988, **173**(1):147-154.

145. Chalkley RJ, Hansen KC, Baldwin MA: **Bioinformatic methods to exploit mass spectrometric data for proteomic applications**. *Methods Enzymol* 2005, **402**:289-312.

146. Patterson SD: **Data analysis--the Achilles heel of proteomics**. *Nat Biotechnol* 2003, **21**(3):221-222.

147. Burlingame AL: **Toward deciphering the knowledge encrypted in large datasets**. *Mol Cell Proteomics* 2003, **2**(7):425.

148. Domon B, Aebersold R: **Challenges and opportunities in proteomics data analysis**. *Mol Cell Proteomics* 2006, **5**(10):1921-1926.

149. Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR, 3rd: **Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility**. *Anal Chem* 2003, **75**(10):2470-2477.

150. Beer I, Barnea E, Ziv T, Admon A: **Improving large-scale proteomics by clustering of mass spectrometry data**. *Proteomics* 2004, **4**(4):950-960.

151. Tabb DL, Thompson MR, Khalsa-Moyers G, VerBerkmoes NC, McDonald WH: **MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra**. *J Am Soc Mass Spectrom* 2005, **16**(8):1250-1261.

152. Gentzel M, Kocher T, Ponnusamy S, Wilm M: **Preprocessing of tandem mass spectrometric data to support automatic protein identification**. *Proteomics* 2003, **3**(8):1597-1610.

153. Flikka K, Meukens J, Helsens K, Vandekerckhove J, Eidhammer I, Gevaert K, Martens L: **Implementation and application of a versatile clustering tool for tandem mass spectrometry data**. *Proteomics* 2007, **7**(18):3245-3258.

154. Zhang X, Hines W, Adamec J, Asara JM, Naylor S, Regnier FE: **An automated method for the analysis of stable isotope labeling data in proteomics**. *J Am Soc Mass Spectrom* 2005, **16**(7):1181-1191.

155. Yague J, Paradela A, Ramos M, Ogueta S, Marina A, Barahona F, Lopez de Castro JA, Vazquez J: **Peptide rearrangement during quadrupole ion trap fragmentation: added complexity to MS/MS spectra**. *Anal Chem* 2003, **75**(6):1524-1535.

156. Mujezinovic N, Raidl G, Hutchins JR, Peters JM, Mechtler K, Eisenhaber F: **Cleaning of raw peptide MS/MS spectra: improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise**. *Proteomics* 2006, **6**(19):5117-5131.

157. Breci LA, Tabb DL, Yates JR, 3rd, Wysocki VH: **Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra**. *Anal Chem* 2003, **75**(9):1963-1971.

158. Tabb DL, Smith LL, Breci LA, Wysocki VH, Lin D, Yates JR, 3rd: **Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides**. *Anal Chem* 2003, **75**(5):1155-1163.

159. Zhang Z: **Prediction of low-energy collision-induced dissociation spectra of peptides**. *Anal Chem* 2004, **76**(14):3908-3922.

160. Hoopmann MR, Finney GL, MacCoss MJ: **High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry**. *Anal Chem* 2007, **79**(15):5620-5632.

161. Sadygov RG, Eng J, Durr E, Saraf A, McDonald H, MacCoss MJ, Yates JR, 3rd: **Code developments to improve the efficiency of automated MS/MS spectra interpretation**. *J Proteome Res* 2002, **1**(3):211-215.

162. Colinge J, Magnin J, Dessingy T, Giron M, Masselot A: **Improved peptide charge state assignment**. *Proteomics* 2003, **3**(8):1434-1440.

163. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R: **Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides**. *Mol Cell Proteomics* 2006, **5**(4):652-670.

164. Flikka K, Martens L, Vandekerckhove J, Gevaert K, Eidhammer I: **Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering**. *Proteomics* 2006, **6**(7):2086-2094.

165. Bern M, Goldberg D, McDonald WH, Yates JR, 3rd: **Automatic quality assessment of peptide tandem mass spectra**. *Bioinformatics* 2004, **20 Suppl 1**:i49-54.

166. Moore RE, Young MK, Lee TD: **Method for screening peptide fragment ion mass spectra prior to database searching**. *J Am Soc Mass Spectrom* 2000, **11**(5):422-426.

167. Xu M, Geer LY, Bryant SH, Roth JS, Kowalak JA, Maynard DM, Markey SP: **Assessing data quality of peptide mass spectra obtained by quadrupole ion trap mass spectrometry**. *J Proteome Res* 2005, **4**(2):300-305.

168. Savitski MM, Nielsen ML, Zubarev RA: **New Data Base-independent, Sequence Tag-based Scoring of Peptide MS/MS Data Validates Mowse Scores, Recovers Below Threshold Data, Singles Out Modified Peptides, and Assesses the Quality of MS/MS Techniques**. *Mol Cell Proteomics* 2005, **4**(8):1180-1188.

169. Junqueira M, Spirin V, Santana Balbuena T, Waridel P, Surendranath V, Kryukov G, Adzhubei I, Thomas H, Sunyaev S, Shevchenko A: **Separating the wheat from the chaff: unbiased filtering of background tandem mass spectra improves protein identification**. *J Proteome Res* 2008, **7**(8):3382-3395.

170. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data**. *Electrophoresis* 1999, **20**(18):3551-3567.

171. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra**. *Bioinformatics* 2004, **20**(9):1466-1467.

172. Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D: **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database**. *Anal Chem* 1995, **67**(8):1426-1436.

173. Zhang N, Aebersold R, Schwikowski B: **ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data**. *Proteomics* 2002, **2**(10):1406-1412.

174. Colinge J, Masselot A, Giron M, Dessingy T, Magnin J: **OLAV: towards high-throughput tandem mass spectrometry data identification**. *Proteomics* 2003, **3**(8):1454-1463.

175. Matthiesen R, Trelle MB, Hojrup P, Bunkenborg J, Jensen ON: **VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins**. *J Proteome Res* 2005, **4**(6):2338-2347.

176. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm**. *J Proteome Res* 2004, **3**(5):958-964.

177. Johnson RS, Taylor JA: **Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry**. *Mol Biotechnol* 2002, **22**(3):301-315.

178. Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA: **De novo peptide sequencing and identification with precision mass spectrometry**. *J Proteome Res* 2007, **6**(1):114-123.

179. Frank A, Pevzner P: **PepNovo: de novo peptide sequencing via probabilistic network modeling**. *Anal Chem* 2005, **77**(4):964-973.

180. Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA: **Proteomics-grade de novo sequencing approach**. *J Proteome Res* 2005, **4**(6):2348-2354.

181. Spengler B: **De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry**. *J Am Soc Mass Spectrom* 2004, **15**(5):703-714.

182. Grossmann J, Roos FF, Cieliebak M, Liptak Z, Mathis LK, Muller M, Gruissem W, Baginsky S: **AUDENS: a tool for automated peptide de novo sequencing**. *J Proteome Res* 2005, **4**(5):1768-1774.

183. Mann M, Wilm M: **Error-tolerant identification of peptides in sequence databases by peptide sequence tags**. *Anal Chem* 1994, **66**(24):4390-4399.

184. Tabb DL, Saraf A, Yates JR, 3rd: **GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model**. *Anal Chem* 2003, **75**(23):6415-6421.

185. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V: **InsPecT: identification of posttranslationally modified peptides from tandem mass spectra**. *Anal Chem* 2005, **77**(14):4626-4639.

186. Nesvizhskii AI, Vitek O, Aebersold R: **Analysis and validation of proteomic data generated by tandem mass spectrometry**. *Nat Methods* 2007, **4**(10):787-797.

187. Johnson RS, Davis MT, Taylor JA, Patterson SD: **Informatics for protein identification by mass spectrometry**. *Methods* 2005, **35**(3):223-236.

188. Nesvizhskii AI, Aebersold R: **Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS**. *Drug Discov Today* 2004, **9**(4):173-181.

189. Venable JD, Yates JR, 3rd: **Impact of ion trap tandem mass spectra variability on the identification of peptides**. *Anal Chem* 2004, **76**(10):2928-2937.

190. Gibbons FD, Elias JE, Gygi SP, Roth FP: **SILVER helps assign peptides to tandem mass spectra using intensity-based scoring**. *J Am Soc Mass Spectrom* 2004, **15**(6):910-912.

191. Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP: **Intensity-based protein identification by machine learning from a library of tandem mass spectra**. *Nat Biotechnol* 2004, **22**(2):214-219.

192. Norbeck AD, Monroe ME, Adkins JN, Anderson KK, Daly DS, Smith RD: **The utility of accurate mass and LC elution time information in the analysis of complex proteomes**. *J Am Soc Mass Spectrom* 2005, **16**(8):1239-1249.

193. Resing KA, Meyer-Arendt K, Mendoza AM, Aveline-Wolf LD, Jonscher KR, Pierce KG, Old WM, Cheung HT, Russell S, Wattawa JL *et al*: **Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics**. *Anal Chem* 2004, **76**(13):3556-3568.

194. Yen CY, Russell S, Mendoza AM, Meyer-Arendt K, Sun S, Cios KJ, Ahn NG, Resing KA: **Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra**. *Anal Chem* 2006, **78**(4):1071-1084.

195. Narasimhan C, Tabb DL, Verberkmoes NC, Thompson MR, Hettich RL, Uberbacher EC: **MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence**. *Anal Chem* 2005, **77**(23):7581-7593.

196. Havilio M, Haddad Y, Smilansky Z: **Intensity-based statistical scorer for tandem mass spectrometry**. *Anal Chem* 2003, **75**(3):435-444.

197. Pieper R, Gatlin CL, Makusky AJ, Russo PS, Schatz CR, Miller SS, Su Q, McGrath AM, Estock MA, Parmar PP *et al*: **The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins**. *Proteomics* 2003, **3**(7):1345-1364.

198. Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD: **Characterization of the low molecular weight human serum proteome**. *Mol Cell Proteomics* 2003, **2**(10):1096-1103.

199. Sun S, Meyer-Arendt K, Eichelberger B, Brown R, Yen CY, Old WM, Pierce K, Cios KJ, Ahn NG, Resing KA: **Improved validation of peptide MS/MS assignments using spectral intensity prediction**. *Mol Cell Proteomics* 2007, **6**(1):1-17.

200. Zhang Z: **Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges**. *Anal Chem* 2005, **77**(19):6364-6373.

201. Wysocki VH, Tsaprailis G, Smith LL, Breci LA: **Mobile and localized protons: a framework for understanding peptide dissociation**. *J Mass Spectrom* 2000, **35**(12):1399-1406.

202. Paizs B, Suhai S: **Fragmentation pathways of protonated peptides**. *Mass Spectrom Rev* 2005, **24**(4):508-548.

203. Moore RE, Young MK, Lee TD: **Qscore: an algorithm for evaluating SEQUEST database search results**. *J Am Soc Mass Spectrom* 2002, **13**(4):378-386.

204. Han DK, Eng J, Zhou H, Aebersold R: **Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry**. *Nat Biotechnol* 2001, **19**(10):946-951.

205. Tabb DL, McDonald WH, Yates JR, 3rd: **DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics**. *J Proteome Res* 2002, **1**(1):21-26.

206. Eddes JS, Kapp EA, Frecklington DF, Connolly LM, Layton MJ, Moritz RL, Simpson RJ: **CHOMPER: a bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies**. *Proteomics* 2002, **2**(9):1097-1103.

207. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search**. *Anal Chem* 2002, **74**(20):5383-5392.

208. MacCoss MJ, Wu CC, Yates JR, 3rd: **Probability-based validation of protein identifications using a modified SEQUEST algorithm**. *Anal Chem* 2002, **74**(21):5593-5599.

209. Anderson DC, Li W, Payan DG, Noble WS: **A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores**. *J Proteome Res* 2003, **2**(2):137-146.

210. Fenyo D, Beavis RC: **A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes**. *Anal Chem* 2003, **75**(4):768-774.

211. Elias JE, Gygi SP: **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry**. *Nat Methods* 2007, **4**(3):207-214.

212. Qian WJ, Liu T, Monroe ME, Strittmatter EF, Jacobs JM, Kangas LJ, Petritis K, Camp DG, 2nd, Smith RD: **Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome**. *J Proteome Res* 2005, **4**(1):53-62.

213. Strittmatter EF, Kangas LJ, Petritis K, Mottaz HM, Anderson GA, Shen Y, Jacobs JM, Camp DG, 2nd, Smith RD: **Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry**. *J Proteome Res* 2004, **3**(4):760-769.

214. Cargile BJ, Bundy JL, Freeman TW, Stephenson JL, Jr.: **Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification**. *J Proteome Res* 2004, **3**(1):112-119.

215. Nesvizhskii AI, Keller A, Kolker E, Aebersold R: **A statistical model for identifying proteins by tandem mass spectrometry**. *Anal Chem* 2003, **75**(17):4646-4658.

216. Shevchenko A, Chernushevich I, Ens W, Standing KG, Thomson B, Wilm M, Mann M: **Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer**. *Rapid Commun Mass Spectrom* 1997, **11**(9):1015-1024.

217. Taylor JA, Johnson RS: **Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry**. *Anal Chem* 2001, **73**(11):2594-2604.

218. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

219. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison**. *Proc Natl Acad Sci U S A* 1988, **85**(8):2444-2448.

220. Shevchenko A, Sunyaev S, Loboda A, Bork P, Ens W, Standing KG: **Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching**. *Anal Chem* 2001, **73**(9):1917-1926.

221. Huang L, Jacob RJ, Pegg SC, Baldwin MA, Wang CC, Burlingame AL, Babbitt PC: **Functional assignment of the 20 S proteasome from Trypanosoma brucei using mass spectrometry and new bioinformatics approaches**. *J Biol Chem* 2001, **276**(30):28327-28339.

222. Taylor JA, Johnson RS: **Sequence database searches via de novo peptide sequencing by tandem mass spectrometry**. *Rapid Commun Mass Spectrom* 1997, **11**(9):1067-1075.

223. Mackey AJ, Haystead TA, Pearson WR: **Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences**. *Mol Cell Proteomics* 2002, **1**(2):139-147.

224. Shevchenko A, Sunyaev S, Liska A, Bork P: **Nanoelectrospray tandem mass spectrometry and sequence similarity searching for identification of proteins from organisms with unknown genomes**. *Methods Mol Biol* 2003, **211**:221-234.

225. Habermann B, Oegema J, Sunyaev S, Shevchenko A: **The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches**. *Mol Cell Proteomics* 2004, **3**(3):238-249.

226. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M: **Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein**. *Mol Cell Proteomics* 2005, **4**(9):1265-1272.

227. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B: **Quantitative mass spectrometry in proteomics: a critical review**. *Anal Bioanal Chem* 2007, **389**(4):1017-1031.

228. Oda Y, Huang K, Cross FR, Cowburn D, Chait BT: **Accurate quantitation of protein expression and site-specific phosphorylation**. *Proc Natl Acad Sci U S A* 1999, **96**(12):6591-6596.

229. Washburn MP, Ulaszek R, Deciu C, Schieltz DM, Yates JR, 3rd: **Analysis of quantitative proteomic data generated via multidimensional protein identification technology**. *Anal Chem* 2002, **74**(7):1650-1657.

230. Conrads TP, Alving K, Veenstra TD, Belov ME, Anderson GA, Anderson DJ, Lipton MS, Pasa-Tolic L, Udseth HR, Chrisler WB *et al*: **Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling**. *Anal Chem* 2001, **73**(9):2132-2139.

231. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics**. *Mol Cell Proteomics* 2002, **1**(5):376-386.

232. Everley PA, Krijgsveld J, Zetter BR, Gygi SP: **Quantitative cancer proteomics: stable isotope labeling with amino acids in cell culture (SILAC) as a tool for prostate cancer research**. *Mol Cell Proteomics* 2004, **3**(7):729-735.

233. Ibarrola N, Molina H, Iwahori A, Pandey A: **A novel proteomic approach for specific identification of tyrosine kinase substrates using [13C]tyrosine**. *J Biol Chem* 2004, **279**(16):15805-15813.

234. Ong SE, Foster LJ, Mann M: **Mass spectrometric-based approaches in quantitative proteomics**. *Methods* 2003, **29**(2):124-130.

235. Yan W, Chen SS: **Mass spectrometry-based quantitative proteomic profiling**. *Brief Funct Genomic Proteomic* 2005, **4**(1):27-38.

236. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. *Nat Biotechnol* 1999, **17**(10):994-999.

237. Oda Y, Owa T, Sato T, Boucher B, Daniels S, Yamanaka H, Shinohara Y, Yokoi A, Kuromitsu J, Nagasu T: **Quantitative chemical proteomics for identifying candidate drug targets**. *Anal Chem* 2003, **75**(9):2159-2165.

238. Hansen KC, Schmitt-Ulms G, Chalkley RJ, Hirsch J, Baldwin MA, Burlingame AL: **Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography**. *Mol Cell Proteomics* 2003, **2**(5):299-314.

239. Li J, Steen H, Gygi SP: **Protein profiling with cleavable isotope-coded affinity tag (cICAT) reagents: the yeast salinity stress response**. *Mol Cell Proteomics* 2003, **2**(11):1198-1204.

240. Yi EC, Li XJ, Cooke K, Lee H, Raught B, Page A, Aneliunas V, Hieter P, Goodlett DR, Aebersold R: **Increased quantitative proteome coverage with (13)C/(12)C-based, acid-cleavable isotope-coded affinity tag reagent and modified data acquisition scheme**. *Proteomics* 2005, **5**(2):380-387.

241. Ranish JA, Yi EC, Leslie DM, Purvine SO, Goodlett DR, Eng J, Aebersold R: **The study of macromolecular complexes by quantitative proteomics**. *Nat Genet* 2003, **33**(3):349-355.

242. Shiio Y, Aebersold R: **Quantitative proteome analysis using isotope-coded affinity tags and mass spectrometry**. *Nat Protoc* 2006, **1**(1):139-145.

243. Schmidt A, Kellermann J, Lottspeich F: **A novel strategy for quantitative proteomics using isotope-coded protein labels**. *Proteomics* 2005, **5**(1):4-15.

244. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C: **Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS**. *Anal Chem* 2003, **75**(8):1895-1904.

245. Ji J, Chakraborty A, Geng M, Zhang X, Amini A, Bina M, Regnier F: **Strategy for qualitative and quantitative analysis in proteomics based on signature peptides**. *J Chromatogr B Biomed Sci Appl* 2000, **745**(1):197-210.

246. Che FY, Fricker LD: **Quantitation of neuropeptides in Cpe(fat)/Cpe(fat) mice using differential isotopic tags and mass spectrometry**. *Anal Chem* 2002, **74**(13):3190-3198.

247. Zhang X, Jin QK, Carr SA, Annan RS: **N-Terminal peptide labeling strategy for incorporation of isotopic tags: a method for the determination of site-specific absolute phosphorylation stoichiometry**. *Rapid Commun Mass Spectrom* 2002, **16**(24):2325-2332.

248. Hsu JL, Huang SY, Chow NH, Chen SH: **Stable-isotope dimethyl labeling for quantitative proteomics**. *Anal Chem* 2003, **75**(24):6843-6852.

249. Ji C, Guo N, Li L: **Differential dimethyl labeling of N-termini of peptides after guanidination for proteome analysis**. *J Proteome Res* 2005, **4**(6):2099-2108.

250. Hsu JL, Huang SY, Chen SH: **Dimethyl multiplexed labeling combined with microcolumn separation and MS analysis for time course study in proteomics**. *Electrophoresis* 2006, **27**(18):3652-3660.

251. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S *et al*: **Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents**. *Mol Cell Proteomics* 2004, **3**(12):1154-1169.

252. Goodlett DR, Keller A, Watts JD, Newitt R, Yi EC, Purvine S, Eng JK, von Haller P, Aebersold R, Kolker E: **Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation**. *Rapid Commun Mass Spectrom* 2001, **15**(14):1214-1221.

253. Brill LM, Salomon AR, Ficarro SB, Mukherji M, Stettler-Gill M, Peters EC: **Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry**. *Anal Chem* 2004, **76**(10):2763-2772.

254. Goshe MB, Conrads TP, Panisko EA, Angell NH, Veenstra TD, Smith RD: **Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses**. *Anal Chem* 2001, **73**(11):2578-2586.

255. Goshe MB, Veenstra TD, Panisko EA, Conrads TP, Angell NH, Smith RD: **Phosphoprotein isotope-coded affinity tags: application to the enrichment and identification of low-abundance phosphoproteins**. *Anal Chem* 2002, **74**(3):607-616.

256. Tao WA, Wollscheid B, O'Brien R, Eng JK, Li XJ, Bodenmiller B, Watts JD, Hood L, Aebersold R: **Quantitative phosphoproteome analysis using a dendrimer conjugation chemistry and tandem mass spectrometry**. *Nat Methods* 2005, **2**(8):591-598.

257. Riggs L, Seeley EH, Regnier FE: **Quantification of phosphoproteins with global internal standard technology**. *J Chromatogr B Analyt Technol Biomed Life Sci* 2005, **817**(1):89-96.

258. Zhang H, Li XJ, Martin DB, Aebersold R: **Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry**. *Nat Biotechnol* 2003, **21**(6):660-666.

259. Yao X, Freas A, Ramirez J, Demirev PA, Fenselau C: **Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus**. *Anal Chem* 2001, **73**(13):2836-2842.

260. Heller M, Mattou H, Menzel C, Yao X: **Trypsin catalyzed 16O-to-18O exchange for comparative proteomics: tandem mass spectrometry comparison using MALDI-TOF, ESI-QTOF, and ESI-ion trap mass spectrometers**. *J Am Soc Mass Spectrom* 2003, **14**(7):704-718.

261. Mirgorodskaia OA, Koz'min Iu P, Titov MI, Savel'eva NV, Korner R, Sonksen C, Miroshnikov AI, Roestorff P: **[MALD-MS in the quantitative analysis of peptides and proteins]**. *Bioorg Khim* 2000, **26**(9):662-671.

262. Schnolzer M, Jedrzejewski P, Lehmann WD: **Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry**. *Electrophoresis* 1996, **17**(5):945-953.

263. Johnson KL, Muddiman DC: **A method for calculating 16O/18O peptide ion ratios for the relative quantification of proteomes**. *J Am Soc Mass Spectrom* 2004, **15**(4):437-445.

264. Ramos-Fernandez A, Lopez-Ferrer D, Vazquez J: **Improved method for differential expression proteomics using trypsin-catalyzed 18O labeling with a correction for labeling efficiency**. *Mol Cell Proteomics* 2007, **6**(7):1274-1286.

265. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP: **Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS**. *Proc Natl Acad Sci U S A* 2003, **100**(12):6940-6945.

266. Pan S, Zhang H, Rush J, Eng J, Zhang N, Patterson D, Comb MJ, Aebersold R: **High throughput proteome screening for biomarker detection**. *Mol Cell Proteomics* 2005, **4**(2):182-190.

267. Kito K, Ota K, Fujita T, Ito T: **A synthetic protein approach toward accurate mass spectrometric quantification of component stoichiometry of multiprotein complexes**. *J Proteome Res* 2007, **6**(2):792-800.

268. Nanavati D, Gucek M, Milne JL, Subramaniam S, Markey SP: **Stoichiometry and absolute quantification of proteins with mass spectrometry using fluorescent and isotope labeled concatenated peptide standards**. *Mol Cell Proteomics* 2007.

269. Beynon RJ, Doherty MK, Pratt JM, Gaskell SJ: **Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides**. *Nat Methods* 2005, **2**(8):587-589.

270. Kirkpatrick DS, Gerber SA, Gygi SP: **The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications**. *Methods* 2005, **35**(3):265-273.

271. Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM: **Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks**. *Proc Natl Acad Sci U S A* 2007, **104**(14):5860-5865.

272. Bondarenko PV, Chelius D, Shaler TA: **Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry**. *Anal Chem* 2002, **74**(18):4741-4749.

273. Chelius D, Bondarenko PV: **Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry**. *J Proteome Res* 2002, **1**(4):317-323.

274. Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P *et al*: **Quantitative proteomic analysis by accurate mass retention time pairs**. *Anal Chem* 2005, **77**(7):2187-2200.

275. Ono M, Shitashige M, Honda K, Isobe T, Kuwabara H, Matsuzuki H, Hirohashi S, Yamada T: **Label-free quantitative proteomics using large peptide data sets generated by nanoflow liquid chromatography and mass spectrometry**. *Mol Cell Proteomics* 2006, **5**(7):1338-1347.

276. Wiener MC, Sachs JR, Deyanova EG, Yates NA: **Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures**. *Anal Chem* 2004, **76**(20):6085-6096.

277. Liu H, Sadygov RG, Yates JR, 3rd: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics**. *Anal Chem* 2004, **76**(14):4193-4201.

278. Rappsilber J, Ryder U, Lamond AI, Mann M: **Large-scale proteomic analysis of the human spliceosome**. *Genome Res* 2002, **12**(8):1231-1245.

279. Heller M, Schlappritzi E, Stalder D, Nuoffer JM, Haeberli A: **Compositional protein analysis of high density lipoproteins in hypercholesterolemia by shotgun LC-MS/MS and probabilistic peptide scoring**. *Mol Cell Proteomics* 2007, **6**(6):1059-1072.

280. Ahn NG, Shabb JB, Old WM, Resing KA: **Achieving in-depth proteomics profiling by mass spectrometry**. *ACS Chem Biol* 2007, **2**(1):39-52.

281. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG: **Comparison of label free methods for quantifying human proteins by shotgun proteomics**. *Mol Cell Proteomics* 2005.

282. Wang P, Tang H, Fitzgibbon MP, McIntosh M, Coram M, Zhang H, Yi E, Aebersold R: **A statistical method for chromatographic alignment of LC-MS data**. *Biostatistics* 2007, **8**(2):357-367.

283. Jaitly N, Monroe ME, Petyuk VA, Clauss TR, Adkins JN, Smith RD: **Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline**. *Anal Chem* 2006, **78**(21):7397-7409.

284. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, May D, Eng J, Fang R, Lin C *et al*: **A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS**. *Bioinformatics* 2006, **22**(15):1902-1909.

285. Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA, Carr SA: **PEPPeR, a platform for experimental proteomic pattern recognition**. *Mol Cell Proteomics* 2006, **5**(10):1927-1941.

286. Bonneil E, Tessier S, Carrier A, Thibault P: **Multiplex multidimensional nanoLC-MS system for targeted proteomic analyses**. *Electrophoresis* 2005, **26**(24):4575-4589.

287. Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ: **Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition**. *Mol Cell Proteomics* 2006, **5**(1):144-156.

288. Gilchrist A, Au CE, Hiding J, Bell AW, Fernandez-Rodriguez J, Lesimple S, Nagaya H, Roy L, Gosline SJ, Hallett M *et al*: **Quantitative proteomics analysis of the secretory pathway**. *Cell* 2006, **127**(6):1265-1281.

289. Havlis J, Shevchenko A: **Absolute quantification of proteins in solutions and in polyacrylamide gels by mass spectrometry**. *Anal Chem* 2004, **76**(11):3029-3036.

290. Cunningham LW, Jr.: **Molecular-kinetic properties of crystalline diisopropyl phosphoryl trypsin**. *J Biol Chem* 1954, **211**(1):13-19.

291. Jouni ZE, Zamora J, Snyder M, Montfort WR, Weichsel A, Wells MA: **alpha-cyclodextrin extracts diacylglycerol from insect high density lipoproteins**. *J Lipid Res* 2000, **41**(6):933-939.

292. Thomas H, Havlis J, Peychl J, Shevchenko A: **Dried-droplet probe preparation on AnchorChip targets for navigating the acquisition of matrix-assisted laser desorption/ionization time-of-flight spectra by fluorescence of matrix/analyte crystals**. *Rapid Commun Mass Spectrom* 2004, **18**(9):923-930.

293. Mirgorodskaya OA, Kozmin YP, Titov MI, Korner R, Sonksen CP, Roepstorff P: **Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards**. *Rapid Commun Mass Spectrom* 2000, **14**(14):1226-1232.

294. Stewart, II, Thomson T, Figeys D: **18O labeling: a tool for proteomics**. *Rapid Commun Mass Spectrom* 2001, **15**(24):2456-2465.

295. Shevchenko A: **Evaluation of the efficiency of in-gel digestion of proteins by peptide isotopic labeling and MALDI mass spectrometry**. *Anal Biochem* 2001, **296**(2):279-283.

296. Regnier FE, Riggs L, Zhang R, Xiong L, Liu P, Chakraborty A, Seeley E, Sioma C, Thompson RA: **Comparative proteomics based on stable isotope labeling and affinity selection**. *J Mass Spectrom* 2002, **37**(2):133-145.

297. Zhu YF, Lee KL, Tang K, Allman SL, Taranenko NI, Chen CH: **Revisit of MALDI for small proteins**. *Rapid Commun Mass Spectrom* 1995, **9**(13):1315-1320.

298. Chrambach A, Rodbard D: **Polyacrylamide gel electrophoresis**. *Science* 1971, **172**(982):440-451.

299. Holmes DL, Stellwagen NC: **Estimation of polyacrylamide gel pore size from Ferguson plots of linear DNA fragments. II. Comparison of gels with different crosslinker concentrations, added agarose and added linear polyacrylamide**. *Electrophoresis* 1991, **12**(9):612-619.

300. Sarbolouki MN, Mahnam K, Rafiee-Pour HA: **Determination of pore/protein size via electrophoresis and slit sieve model**. *Electrophoresis* 2004, **25**(17):2907-2911.

301. Farmer WH, Yuan ZY: **A continuous fluorescent assay for measuring protease activity using natural protein substrate**. *Anal Biochem* 1991, **197**(2):347-352.

302. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A: **The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data**. *Mol Cell Proteomics* 2004, **3**(6):531-533.

303. Altschul SF, Gish W: **Local alignment statistics**. *Methods Enzymol* 1996, **266**:460-480.

304. Gruss OJ, Vernos I: **The mechanism of spindle assembly: functions of Ran and its target TPX2**. *J Cell Biol* 2004, **166**(7):949-955.

305. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T *et al*: **A map of the interactome network of the metazoan C. elegans**. *Science* 2004, **303**(5657):540-543.

306. Camilleri C, Azimzadeh J, Pastuglia M, Bellini C, Grandjean O, Bouchez D: **The Arabidopsis TONNEAU2 gene encodes a putative novel protein phosphatase 2A regulatory subunit essential for the control of the cortical cytoskeleton**. *Plant Cell* 2002, **14**(4):833-845.

307. Turkova J, Vohnik S, Helusova S, Benes MJ, Ticha M: **Galactosylation as a tool for the stabilization and immobilization of proteins**. *J Chromatogr* 1992, **597**(1-2):19-27.

308. Morand P, Biellmann JF: **Modification of alpha-amylase from Bacillus licheniformis by the polyaldehyde derived from beta-cyclodextrine and alpha-amylase thermostability**. *FEBS Lett* 1991, **289**(2):148-150.

309. Nureddin A, Inagami T: **Chemical modification of amino groups and guanidino groups of trypsin. Preparation of stable and soluble derivatives**. *Biochem J* 1975, **147**(1):71-81.

310. Hartree EF: **Determination of protein: a modification of the Lowry method that gives a linear photometric response**. *Anal Biochem* 1972, **48**(2):422-427.

311. Dubois M, Gilles K, Hamilton JK, Rebers PA, Smith F: **A colorimetric method for the determination of sugars**. *Nature* 1951, **168**(4265):167.

312. Habeeb AF: **Determination of free amino groups in proteins by trinitrobenzenesulfonic acid**. *Anal Biochem* 1966, **14**(3):328-336.

313. Robertson EF, Dannelly HK, Malloy PJ, Reeves HC: **Rapid isoelectric focusing in a vertical polyacrylamide minigel system**. *Anal Biochem* 1987, **167**(2):290-294.

314. Schagger H, von Jagow G: **Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa**. *Anal Biochem* 1987, **166**(2):368-379.

315. Frank A, Tanner S, Bafna V, Pevzner P: **Peptide sequence tags for fast database search in mass-spectrometry**. *J Proteome Res* 2005, **4**(4):1287-1295.

PUBLICATIONS

Ozlu N, Srayko M, Kinoshita K, Habermann B, O'Toole E T, Muller-Reichert T, Schmalz N, Desai A, Hyman AA: An essential function of the C. elegans ortholog of TPX2 is to localize activated aurora A kinase to mitotic spindles. ***Dev Cell 2005*, 9(2):237-248.**

Wielsch N, Thomas H, Surendranath V, Waridel P, Frank A, Pevzner P, Shevchenko A: Rapid validation of protein identifications with the borderline statistical confidence via de novo sequencing and MS BLAST searches. ***J Proteome Res* 2006, 5(9):2448-2456.**

Šebela M, Stosova T, Havliš J, Wielsch N, Thomas H, Zdrahal Z, Shevchenko A: Thermostable trypsin conjugates for high-throughput proteomics: synthesis and performance evaluation. ***Proteomics* 2006, 6(10):2959-2963.**

Schlaitz AL, Srayko M, Dammermann A, Quintin S, Wielsch N, MacLeod I, de Robillard Q, Zinke A, Yates JR, 3rd, Muller-Reichert T et al: The C. elegans RSA complex localizes protein phosphatase 2A to centrosomes and regulates mitotic spindle assembly. ***Cell* 2007, 128(1):115-127.**

ERKLÄRUNG ENTSPRECHEND § 5.5 DER PROMOTIONSORDNUNG

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die Dissertation wurde im Zeitraum von Oktober 2003 bis April 2008 am Max Plank Institut für Molekulare Zellbiologie und Genetik, Abteilung Biologische Massenspektrometrie, Dresden unter der wissenschaftlichen Betreuung von Dr. Andrej Shevchenko angefertigt.

Meine Person betreffend erkläre ich hiermit, dass keine früheren erfolglosen Promotionsverfahren stattgefunden haben.

Ich erkenne die Promotionsordnung der Fakultät für Mathematik und Naturwissenschaften, Technische Universität Dresden an.

DECLARATION ACCORDING TO § 5.5 OF THE DOCTORATE REGULATIONS

Herein, I declare that I have produced this manuscript without the prohibited assistance of third parties and without making use of aids other then those specified; notions taken over directly or indirectly from other sources have been identified as such. This manuscript has not been presented in identical or similar form to any German or foreign examination board. Experimental work performed by collaborators is indicated as such.

The thesis work was conducted from October 2003 to March 2008 under the supervision of Dr. Andrej Shevchenko at the Max Plank Institute of Molecular Cell Biology and Genetics, Dresden in the biological mass spectrometry laboratory.

I declare that I recognize the doctorate regulations of the Faculty of Sciences of the Technische Universität Dresden.